THE CONTRIBUTION OF MOSAIC MUTATION TO AUTISM SPECTRUM DISORDER

by
Donald Freed

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

January, 2017

Abstract

Background

Genetic variation arises as the result of both spontaneous and artificial processes. While genetic variation and natural selection are the only tools of evolution, genetic variation is also known to also cause disease, including inherited disease and cancer. Here we explore the consequences of human genetic variation first as a tool for testing identity from a mixture of DNA and second as a contributor to autism spectrum disorder.

Methods

One of the most comprehensive datasets of human genetic variation across all human populations is the 1000 Genomes Project. We use genetic variants from the 1000 Genomes Project to identify polymorphic loci in extant human populations using custom computational tools. For our studies of mosaic mutation in individuals with autism, we use publicly available data and a combination of publically available and custom software.

Results

Through our analysis of publically available data, we find genomic loci that may be used for identity testing across many human populations. In addition we show that mosaic genetic variation detectable in blood contribute significantly to autism spectrum disorder while mosaic genetic variants that are unique to affected tissues are not frequently detectable from bulk sequence data.

Conclusions

Our results have implications for the fields of identity testing and disease genetics. Using the genomic loci we identify, it is possible to develop assays to identify DNA from an

individual within a mixture even if that individual's DNA makes up less than one-millionth of the total DNA in the mixture.

Our identification of a contribution of mosaic mutations to disease has implications for our understanding of heritability. In classic measurements of heritability, identical twins are used as individuals of constant genetic background and any phenotypic differences between identical twins are assumed to come from the "environment". While genetic variation unique to a single individual within a twin pair has been identified, our results are the first to indicate that these mutations play a role in the phenotypic differences between identical twins. Our results also have implications for the field of genetic counseling. The parents of probands with high-confidence mosaic mutations may be less likely to have additional children with an autism diagnosis compared to parents with children with ASD overall.

Advisor: Jonathan Pevsner

Reader: Robert Scharpf

Acknowledgements

First and foremost I would like to thank the patients and their families, without whom this research would not be possible. I am grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). I thank Laurence Frelin and Alexis Norris for technical advice and support in sample preparation and validation experiments. I thank Eric Stevens for assistance in writing the review article on somatic mosaicism in the human genome. I thank Robert Scharpf, Ben Langmead, Jeremy Nathans, and Ingo Ruczinski for helpful discussions over the course of my graduate carrer. Dr. Scharpf especially helped in generating statistical models that were used for data analysis. I thank Roxann Ashworth and Jocelyn Henline of the Johns Hopkins Genetic Resources Core Facility with assistance performing pyrosequencing experiments. Data were obtained from the NIH supported National Database for Autism Research (NDAR). I thank Illumina, Inc. for Platinum Genomes data (http://www.illumina.com/platinumgenomes/).

I would like to thank all the members of the Pevsner lab and my fellow members of the BCMB program for making my graduate career much more enjoyable. Lastly I would like to thank my mentor Jonathan Pevsner for his support, his generosity with his time, and his personal and professional guidance.

Table of Contents

List of Tables

List of Figures

Chapter 1: A Haplotype-Based Method for the Detection of Chimerism from Next-Generation Sequencing Data

1.1 Introduction

The ability to identify individuals solely on the basis of their DNA sequences has revolutionized modern forensics, where blood, semen, or hair found at a crime scene can be conclusively linked to a single individual on the basis of polymorphisms at short tandem repeats (STR) [1]. While two unrelated individuals may share any given genetic locus, it is extremely unlikely that unrelated individuals will share the same alleles across many loci. However, these methods are not sensitive for identifying DNA from particular individuals that are present at low levels in a mixture. The problem of low sensitivity is especially salient when considering individuals who have undergone myeloablative conditioning followed by bone marrow engraftment, usually for the treatment of hematological malignancies. After treatment, the peripheral blood leukocytes of these individuals will originate from the donor, however the observation of peripheral blood leukocytes originating from the patient is indicative of cancer recurrence [2].

Given the low sensitivity of STR-based methods for detection of patient blood chimerism, alternative techniques may be fruitful. One especially intriguing method is next-generation DNA sequencing, which is capable of generating many reads covering a potential locus at low cost. The error rate of one of the most popular technologies, Illumina, is around 1%, indicating that sequencing of a single SNP that is known to be different between the patient and donor could potentially improve on the STR-based approach. However, a more powerful approach would be to examine loci with haplotypes of SNPs that are different between patient and donor. Assuming errors at the SNPs are independent, the likelihood of a

sequencing read that originates from the donor actually containing the sequence of the patient would be 0.01 to the power of n where n in the number of SNPs present in the patient and absent from the donor that are covered by the sequencing read. While this method can be used to easily distinguish between different individuals at the HLA loci, other loci may be useful for cases where patent and donor have very similar HLA types. To aid in this effort, we identified genomic loci that are likely to contain multiple polymorphisms from two unrelated individuals using the 1000 Genomes data.

1.2 Materials and Methods

Phased variants were downloaded from the 1000 Genomes website (release version 3_20110521; http://www.1000genomes.org; last accessed July 3, 2014). Loci across the genome were then considered for use in the analysis. Loci were considered as candidates if they contained at least 9 variants of at least 9% allele frequency in the CEU, JPT/CHB, and YRI populations. Given two haplotypes, a comparison between these two haplotypes would be considered informative if there were at least two SNP differences between the haplotypes. The genetic diversity of candidate loci was then measured as the probability of a comparison of haplotypes being informative given that two haplotypes were drawn from the population.

1.3 Results

We identified 4,349 loci harboring nine SNPs with allele frequencies of at least 9%. Through our analysis we determined that 7 non-HLA loci were likely to be informative at least 70% of the time in two individuals drawn at random in the population. Importantly, almost all of the identified loci are far enough apart in the genome that they segregate independently in the population indicating that testing of multiple loci could increase sensitivity for chimerism.

1.4 Discussion

Through the use of publically available data and custom computational methods, we identify genomic loci that are likely to be informative when testing for mixtures of DNA from distinct individuals. We imagine that these loci will potentially be valuable for testing of chimerism in individuals who have undergone myeloablative conditioning followed by bone marrow engraftment although more work remains to be done before these loci may be used in the clinic. Especially important will be validation of the diversity of these loci across a population of patient samples.

Sensitive methods for detection of chimerism in mixtures have many applications outside of bone marrow engraftment monitoring. For instance, this method might be used to detect DNA of a suspect at a crime scene, even if the suspect's DNA is present at a low fraction within a mixture. Additionally, a haplotype-based approach may be useful when measuring sample contamination, such as contamination of a sample with a technician's DNA that may occur during sample preparation [3].

In the future, this approach will be aided by the generation of long reads or synthetic long reads from third generation sequencing technologies. Already Illumina sequencing machines are capable of generating paired 300bp reads, suggesting that the size of the genomic loci used in our analysis could be expanded to 500bp or larger. Currently available Pacific Biosciences and Oxford Nanopore sequencing machines are capable of generating 10 kb sequence reads, albeit with much higher rates of errors per base than Illumina sequencers. One can imagine that given the known sequence of the donor and patient, these technologies may allow for the detection of chimerism with ultra-high accuracy from a relatively small number of sequence reads collected in an unbiased manner.

Chapter 2: Somatic Mosaicism in the Human Genome

2.1 Introduction to Somatic Mosaicism

2.1.1 Early Studies of Mosaicism

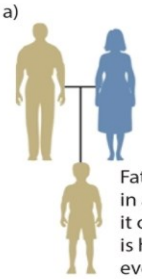
Somatic mosaic mutations are defined as mutations that occur in some cells of the soma of a single individual (Figure 2.1) [4,5]. The mixture of mutation-positive cells with non-mutated cells results in an individual who is a mosaic, or contains different DNA within different cells of his or her body. Mosaic mutations may be present in the germline or soma, however, typically only mutations in the soma have phenotypic consequences or are detectable by currently available genotyping methods. Mosaic mutations in germ cells are usually only discovered when they are inherited by multiple progeny. *De novo* mutations are defined as mutations arising uniquely in a cell that were absent from the cell's parent cell while germline *de novo* mutations are operationally defined as mutations found in all cells of an individual but not detected in that individual's parents (Figure 2.1 d,e) [6]. *De novo* mutations only present in the offspring may occur very early in development; however, this is rare and increasingly sensitive genetic assays are discovering low-level parental mosaicism in supposedly *de novo* cases (Figure 2.1 b) [7,8].

Figure 2.1. Overview of categories of variation.

Inherited mutations are always transmitted through the germline (a), although a parent may also have a somatic mutation (b). In such cases, a child may inherit the variant as a heterozygous mutation with a more severe clinical phenotype. A parent may also have germline mosaicism which may be inherited by progeny (c). *De novo* mutations are operationally defined as genotypes observed in a child but not in either parent. They may originate in a parental germ cell (as may be inferred in a pedigree having multiple affected offspring) (d) or postzygotically (e). Somatic mutation may occur relatively early in development (f) or at any time throughout the lifespan (g), generally affecting fewer cells.

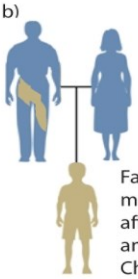
Inherited

a)



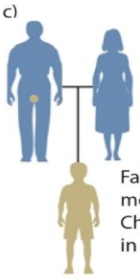
Father has mutation in all cells and transmits it on to his son. Child is heterozygous in every cell.

b)



Father has mosaic mutation that affects germline and somatic cells. Child is heterozygous in every cell.

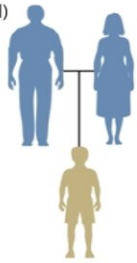
c)



Father has germline mosaic mutation. Child is heterozygous in every cell.

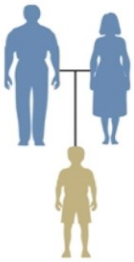
De novo

d)



Father has mutation in a single sperm cell and transmits it to the son. Child is heterozygous in every cell.

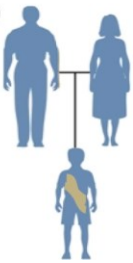
e)



Mutation occurs in zygote within first few cell divisions. Child is heterozygous in every cell.

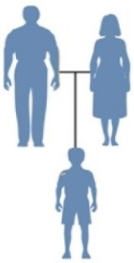
Somatic

f)



Child has mosaic somatic mutation that occurrs early in postzygotic development and is present in a percentage of his cells.

g)



Child has mosaic somatic mutation that occurrs later in development and affects fewer cells (e.g. skin cells,)

The role of somatic genetic changes in human health has been considered at least since 1914 when Theodor Boveri recognized that cancers frequently have abnormal karyotypes [9]. Alfred Knudson built upon the work of Boveri and others and in 1971 described a two-hit model of cancer resulting from both an inherited germline mutation and a later somatic mutation [10]. The model of metastatic cancer occurring as a result of multiple mutations in single cells has remained largely unchanged for over forty years [11,12].

The scientific community was slower to realize the importance of postzygotic mutational events outside of cancer. In the early 1950s, Barbara McClintock demonstrated the phenotypic importance of somatic transposition in *Zea mays* and in 1959 Sir Macfarlane Burnet proposed a role for somatic mutation in disease [13,14]. Nonetheless few studies indicated that somatic mosaic mutations played a role in human health. This changed in the 1970s with the discovery that somatic gene rearrangement creates functional diversity of immunoglobulin and T-cell receptor genes [15-17]. Today, it is known that somatic mutations are ubiquitous [18] and have important roles in cancer [12], aging [19,20], neurodegeneration [21], monogenic disease [22-24], reversion of inherited disease [25-28], and numerous neurocutaneous disorders [29].

2.1.2 Categories of Somatic Variation

Somatic variation has been observed at all genomic scales from point mutations to aneuploidies. At the level of whole chromosomes and large chromosomal segments, complex genomic rearrangements occur somatically (as well as in the germline). The loss or gain of entire chromosomes is thought to be caused by errors in chromosomal segregation during anaphase while non-allelic homologous recombination may cause the loss, gain, or rearrangement of large genomic regions [30,31]. The phenotypic consequences of these

events vary considerably based on the size of the event and the genomic region involved.

In some instances, both copies of a chromosome pair (or of a chromosomal segment) are inherited from one parent, a phenomenon termed uniparental disomy (UPD) [32,33]. UPD may involve two copies from a parent that are identical (uniparental isodisomy) or different (uniparental heterodisomy). Either form may disrupt epigenetically imprinted regions (defined as undergoing differential expression depending on the parent of origin), while uniparental isodisomy may also expose two copies of a recessive mutation. One mechanism for the occurrence of UPD involves trisomic rescue in which an extra (third) copy of a chromosome is rejected, producing a diploid cell line in which one parent's monoploid copy is lost [34]. Frequently, the trisomic rescue is restricted to a fraction of cells in an individual resulting in mosaic trisomy/UPD [35]. UPD may also result from somatic recombination occurring from a reciprocal exchange during mitosis, leading to loss of heterozygosity.

RNA-templated DNA polymerases are another cause of genomic instability. While numerous types of repetitive elements are present in human genomes, only non-long terminal repeat retrotransposons are currently competent for transposition [36]. Successful retrotransposition of these elements is dependent upon functional protein products from long interspersed elements (LINEs). In most somatic tissues, LINEs are epigenetically suppressed; however, these elements escape epigenetic repression during early embryonic development, and their integration into other functional genomic elements occasionally results in disease. One example is one case of choroideremia (OMIM #303100) where a patient was found to have a full L1 repeat inserted into the coding region of the *CHM* gene [37]. In somatic tissues with unusual epigenetic states retrotransposition may also occur [38].

Low complexity regions, including trinucleotide repeats, are scattered throughout the mammalian genome. Trinucleotide repeats can be hypervariable and expansion of some trinucleotide repeats is the cause of nearly 30 disorders [39,40]. The molecular mechanisms underlying expansion or contraction of these regions are complex and cause these regions to have variable length throughout the body of those afflicted with disease [41-47].

Small genetic aberrations may be caused by a number of mechanisms. Polymerase errors may result in nucleotide misincorporation or small insertions or deletions in the germline or soma. Over time, DNA will accumulate numerous lesions and DNA polymerization across these lesions is especially error-prone. DNA lesions may be detected and repaired prior to DNA polymerization, but lesion repair may also create single nucleotide variants, or small insertions or deletions [19,48]. Importantly, the process of transcription creates single-stranded DNA across genes and this single-stranded DNA is especially prone to mutation with a distinct mutational signature [49,50].

In linear mammalian genomes, DNA replication starts at multiple origins with DNA polymerases ε and δ [51,52]. Polymerase ε moves processively 5' to 3' along the genome on the leading strand, moving in the same direction as the replication fork. On the lagging strand replication by polymerase δ also proceeds 5' to 3' but in the opposite direction as the replication fork, causing replication of that strand to be iterative. This process works well for the majority of the genome, but incomplete replication of the lagging strand leads to loss of genetic information at the ends of the chromosome during every replication [53]. This end replication problem is solved in the germline as the ends of chromosomes, telomeres, are protected by repetitive DNA which is synthesized by a dedicated RNA-templated DNA polymerase called telomerase [53]. However, telomerase is not usually expressed in somatic

tissues, likely as a method of protection against malignant transformation, and decreased telomere length is a form of somatic variation.

2.1.3 Mosaicism During Development

A defining characteristic of mosaic mutations is that they occur postzygotically and are inherited by all subsequent cells in their lineage (Figure 2.1). Somatic changes in early development are known to induce an extraordinarily high rate of aneuploidy. 15-20% of clinically recognized pregnancies result in spontaneous abortion, and half of these are attributed to aneuploidy [32]. A review of 36 published studies showed that of 815 human preimplantation embryos, only 177 (22%) were diploid while 73% were mosaic for copy number alterations [54]. In most cases these were diploid-aneuploid mosaic embryos, having one or more diploid cells as well as other cells that were haploid or polyploid for a particular chromosome. Mitotic errors could account for the high rate of chromosomal mosaicism.

Due to the exponential rate of growth during development, somatic mutations must occur early in development to have phenotypic effects over large portions of the body. Indeed, studies of somatic mutation events from single cells collected across the body indicate that mutations that are present in more than approximately 10% of cells in a given tissue will be dispersed throughout the body while mutations present at lower frequencies are tissue specific [55]. Severe somatic mutations which would be embryonic lethal if inherited have an even shorter window during development in which they must occur to be observed in adults [23]. Occurring early in development, these mutations are embryonically lethal; occurring later in development they may have little or no obvious phenotypic effect.

Mutations that alter cellular growth do not necessarily have to occur within such a short developmental window. Inactivating mutations in genes encoding tumor suppressors or

activating mutations in oncogenes may have organism-level phenotypic consequences regardless of when they occur, as evident from their role in cancer. On the other hand, growth retarding mutations, such as inactivating mutations in oncogenes or certain cyclins, are unlikely to have phenotypic effect in adults regardless of when they occur in development as the total number of cells containing the mutation is likely to be small.

Somatic mutations are thought to occur in all cells during replication. On average, 50 mutations occur in microsatellite regions during every mitotic division of a given cell [18]. Mutations in microsatellites and other regions of the genome, assessed by either single-cell or deep sequencing, can then be used to infer cell lineage trees [56]. To date, the most successful lineage tracing experiments have made use of increasingly sophisticated microscopic techniques [57]. However, microscopy-based approaches have practical and technological barriers such as the requirement that non-transgenic cells must be monitored over time. Recent advances in whole genome amplification (WGA) and second-generation sequencing (SGS) offer genetic-based approaches that do not have the same limitations. Already, these techniques have been used to provide a detailed view of the genetics of cancer metastasis [58,59].

2.1.4. Mosaicism Across the Body

By definition, somatic mosaic mutations affect only a subset of cells within an individual (Figure 2.1). This is most easily visible in monogenic mutations affecting pigmentation patterns. While such patterns may be mistaken for stochastic X chromosome inactivation or autoimmune response, somatic mutation is generally localized over a small portion of the body and in many cases occurs along lines of Blaschko [60]. To date, almost all non-cancerous somatic mutations characterized at the molecular level result in visible

abnormalities, usually involving hypertrophy (cellular overgrowth) or abnormal pigmentation [29,60]. Some of our inability to identify mutations that do not result in visible phenotypes is practical; during dissection it is difficult to distinguish affected from unaffected tissue. However, due to the current emphasis on visible phenotypes, few data are available on the extent to which non-visible somatic mutations influence important biological processes.

An important consideration is that somatic mutations occur in varying cell types and tissues as well as different developmental stages. This raises the possibility that a specific mutation may vary in its clinical importance depending on where the mutation occurs across the body. Mutations in *GNAQ* provide an example. Our lab identified p.Arg183Gln mutations in *GNAQ,* encoding the G protein alpha subunit Gαq, as the cause of both Sturge-Weber syndrome (OMIM #185300) and port-wine stain birthmarks (OMIM #163000) [61]. Port-wine stains are non-syndromic vascular abnormalities while the Sturge-Weber syndrome is a severe neurocutaneous disorder, although both conditions likely affect some of the same cell types (e.g. endothelial cells). The milder phenotype of the birthmarks could result from a later developmental origin of the mutation during fetal development or a different cellular lineage harboring the mutation. The identical p.Arg183Gln mutation in *GNAQ*, when occurring in melanocytes during later in life contributes to uveal melanoma, highlighting the importance of both the location and timing of the mutation. p.Arg183Gln mutations in different cell types and developmental stages could have different phenotypic consequences, if any [62].

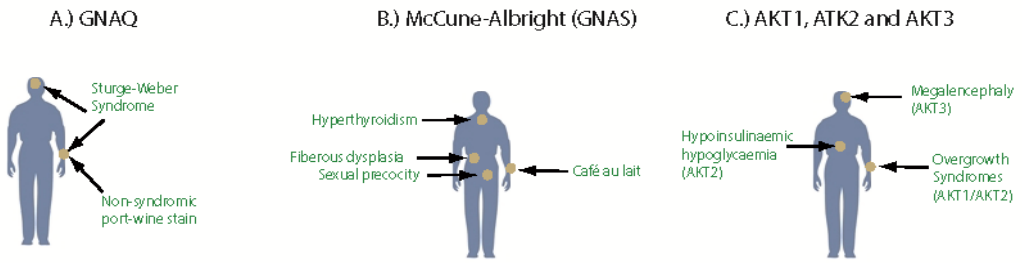
Other mosaic mutations also differ in their clinical importance based on cell or tissue-specific involvement. McCune-Albright syndrome (OMIM #174800) is characterized by increased function of endocrine glands, sexual precocity, café-au-lait macules, and fibrous

dysplasia. These symptoms can vary considerably based, in part, on the extent of the mutation [63]. Like Sturge-Weber syndrome, this disorder is caused by somatic activating mutations in a gene encoding a G protein alpha subunit (*GNAS* encoding Gαs). Expression of this gene highlights another dimension of mosaicism. *GNAS* is expressed biallelically through most of the body, but the maternal allele is imprinted in particular tissues such as the pituitary. The disorders progressive osseous heteroplasia (OMIM #166350) and pseudopseudohypoparathyroidism (OMIM #612463) result from loss of function mutations in the paternal allele of *GNAS* [64].

Somatic mutations in three *AKT* genes also have cell-specific effects [65-67]. Somatic AKT1 mutations are associated with somatic breast cancer, colorectal cancer, and ovarian cancer as well as the Proteus syndrome. The AKT2 gene is expressed selectively in insulin-responsive tissues and mutations are associated with diabetes. Somatic mutations in AKT3 cause Megalencephaly-polymicrogyria-polydactyly-hydrocephalus syndrome 2 (OMIM *611223). Given the localized nature of somatic mutations in AKT discovered to date, it is likely that mutations in these genes occurring outside of vulnerable cell types have few effects. These examples highlight the complex interaction of localized somatic mutation with tissue or cell-specific gene expression and signaling pathways (Figure 2.2).

Figure 2.2. Tissue-specific effects of mutations in *GNAQ*, *GNAS* and *AKT1*, *AKT2*, and *AKT3*.

Constitutively activating mutations in *GNAQ* may lead to either Sturge-Weber syndrome, nonsyndromic port-wine stain, or uveal melanoma (A). Somatic activating mutations in *GNAS* lead to McCune-Albright syndrome which may involve variable hyperthyroidism, *café au lait* macules and sexual precocity (B). Activating mutations in all three of the *AKT* genes cause cellular overgrowth phenotypes with mutations in *AKT2* also implicated in abnormal insulin signaling (C).



Numerous studies have aimed to assess the prevalence of mosaic alterations in tissues of apparently normal individuals. Reanalysis of data from multiple large genome-wide association studies have determined that the number of detectable mosaic events rises sharply after age 50. Furthermore, individuals with increased numbers of mosaic events have higher risk for developing cancer [68,69]. While this measured increase of mosaicism may be due to increased rates mutation rates in elderly individuals, it is much more likely that these events are the result of clonal expansion and positive selection within the stem cell niche or decline in the total number of hematopoietic stem cell progenitors later in life. Notably, increased rates of mosaicism in apparently normal tissues have been linked to poorer prognosis in individuals with ovarian cancer [70].

Studies of twins have demonstrated that post-zygotic mutations may be phenotypically important. Notable examples are monozygotic twins who are discordant for phenotypic sex due to mosaic loss of chromosome Y [71,72]. Numerous examples of monozygotic twins exist where either the presence [73,74] or severity [75] of disease is discordant between twin pairs due to variable proportions of mosaic cells.

Studies of multiple tissues of apparently normal individuals have also found evidence for mosaic events. Analysis of CNVs using hybridization of DNA from multiple tissues of three apparently normal individuals to bacterial artificial chromosome arrays found evidence for six somatic CNVs [76]. Higher resolution examination of a total 33 tissues from six individuals using array comparative genomic hybridization found evidence for 73 high-confidence mosaic CNVs, although a majority of high-confidence events (54/73) were found in one of two particular tissues [77]. It has been noted that induced pluripotent stem cells (iPSC) frequently contain CNVs which may be caused by genomic instability inheritent to

the process of immortalization. Abyzov et al. performed a detailed study of this phenomenon and concluded that almost half of CNVs present in iPSC lines can be found in the parental fibroblasts. Furthermore, they conclude that approximately 30% of all fibroblasts in their sample contain some mosaic CNVs [78].

While experimentation with bulk tissues has shown that somatic mosaicism occurs frequently in normal populations, the combination of DNA from many cells limits the ability of an assay to detect mosaic events unique to single or few cells. As a result, sequencing of single-cells has been recently used to assay mosaicism in normal tissues. These methods have been used to sensitively reexamine conclusions regarding the extent of mosaicism in the brain. Previous reports had indicated that up to 33% of neuroblasts were aneuploid while up 80 retrotransposon insertions occur per neuron [79-82]. Single-cell experiments of the same phenomena have shown that large copy-number variants occur in over 14% of neurons but whole chromosome aneuplodies and retrotransposition events are relatively rare [83-85].

Single-cell studies have also been used to investigate the extent to which mosaicism occurs in early development. It has been known since 1983 that chorionic villus sampling may indicate the presence of a trisomy, while the fetus is diploid without the presence of mosaicism, a condition termed confined placental mosaicism [86-88]. Single-cell studies of young embryos cultured in vitro also demonstrate that chromosomal aneuploidies are common and were found in 83% of tested embryos [89]. While it is likely that many aneuploid embryos are unlikley to result in viable pregnancies, recent advances in prenatal testing allow for the sensitive and specific detection of numerous trisomies by sequencing of circulating fetal DNA from maternal plasma [90].

2.2 Detection of Somatic Mosaicism

2.2.1. Technical Considerations

Almost every type of genetic variation has been implicated as a source of somatic variation including expansion of trinucleotide repeats, point mutation, copy-number loss/gain, uniparental disomy, mitotic recombination, aneuploidy, translocation, and retrotransposition [41-44,47,73,81,84,91-99]. The techniques summarized below vary widely in their ability to detect specific types of somatic variation and more specialized techniques exist for the sensitive detection of some types of variation.

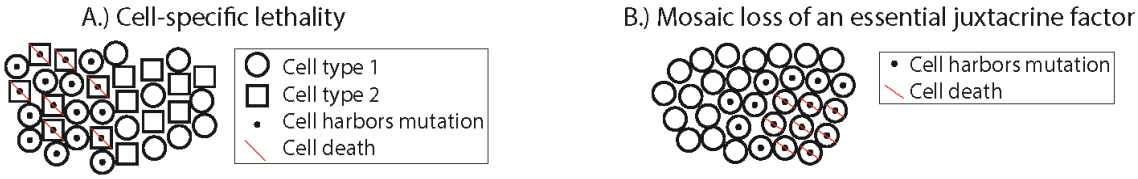
A primary consideration during the analysis of mosaic samples is the purity of the dissection from tissue samples. The presence of normal cells in affected tissue significantly decreases the ability to detect mosaic alterations. This problem can be compounded by the fact that cellular migration during development is prevalent in some tissues. Thus, in a tissue affected by a somatic mutation, two neighboring cells may both be affected if they share a common lineage from the mutated cell. Alternatively, cellular migration could cause neighboring cells to originate from distinct precursors with only one cell affected. Cellular migration can place an important biological constraint on the visible frequency of driver somatic mutations in affected tissues (e.g. in the brain) [6,100,101].

While contamination of normal cells is known to decrease the observed frequency of mosaic mutations, other mechanisms may decrease the detectable fraction of mosaic cells within a sample. Two possibilities are cell-type specific lethality and mosaic absence of essential juxtacrine or paracrine signaling factors. Cellular signaling pathways are known to have cell-type specific effects raising the possibility that a mosaic mutation may be lethal in only one type of cell within a tissue. Furthermore, some paracrine or juxtacrine signaling factors are essential for cell viability [102-104]. Mosaic loss of these factors could result in

affected tissue that is dependent upon surrounding normal tissue for survival, reducing the total number of mutant cells (Figure 2.3).

Figure 2.3. Cell death may reduce the total number of cells harboring somatic mutation.

Mosaic mutations may cause cell-type-specific lethality (A). Mosaic loss of an essential juxtacrine signaling factor may cause localized cell death of cells are not adjacent to unaffected tissues (B).



In Sturge-Weber affected tissues, detected *GNAQ* mutant allele frequencies were between 1 and 18% [61]. Other studies using similar techniques have detected mutant allele frequencies of 1-47%,[66] 3-30% [105], and 3-35% [65] for causative mutations in individuals with Proteus syndrome (OMIM #176920), CLOVES (OMIM #612918), and hemimegalencephaly (e.g. OMIM #611223), respectively. Such relatively low allele frequencies are likely explained by the presence of low proportions of affected cells in a given tissue.

In second-generation sequencing experiments, sequencing and mapping errors are a major concern, as some portions of the genome are known to be prone to false-positive variant calls [106]. Recent improvements in sequencing chemistry have lowered the frequency of sequencing errors. However, biased errors in sequencing are still problematic for the detection of somatic variation, especially when the mutant allele frequency may be close to the technology's inherent error rate. Generally, ultra-high depth sequencing (> 500 reads) of normal and affected tissues will permit detection of these errors. However, exploratory studies generally do not reach this level of depth. It is likely that without validation, these errors are a source of false positives in somatic variation databases. Comparing suspected somatic mutations across multiple tissue types from multiple individuals may be a possible solution to this problem [107].

2.2.2. Cytogenetics

Microscopy-based methods allow for the detection of large mosaic events in single cells. Early cytogenetic methods for identifying extra or fewer chromosomes involved counting condensed metaphase chromosomes under a microscope [108]. Later methods using Giemsa staining and other dyes produced unique chromosomal bands allowing for the

identification of intra- and interchromosomal translocations, duplications, deletions, and large structural rearrangements. However, this method can only resolve aberrations larger than 3-10 Mb [109]. Other methods, such as fluorescent in situ hybridization (FISH), label a specific region of the genome by hybridization of a fluorescent probe allowing for the detection of deletions and some duplications [110]. Variations in this methodology exist using multiple probes of different color to detect several unique fragments at a time (i.e. multicolor FISH). These methods are able to achieve resolutions below 100 kilobases or, in some cases, as few as several kilobases [111]. Potential probe binding to off-target regions is a major consideration in most FISH experiments and adequate controls are required to confirm locus specificity [111]. Variants on classical FISH methods continue to be developed which promise to increase the ability of fluorescent probes to detect small chromosomal abnormalities across increasingly large portions of the genome [111,112]. In combination with high-throughput techniques, these approaches may be used to screen large numbers of cells from a single individual allowing for the detection of low levels of mosaicism.

2.2.3. Genome-Wide Arrays

Comparative genomic hybridization (CGH) is a technique in which fluorophore-labeled DNA from a control and test individual are hybridized to a metaphase reference chromosome [113]. The ratio of fluorescence emission is then measured to allow for the detection of duplication or deletions. A ratio of 1:1 indicates that both samples of DNA carry the same copy number while deviations from this ratio indicate a copy number variant [114].

Two principal array-based techniques that have emerged as alternatives to CGH are array CGH (aCGH) and single nucleotide polymorphism microarrays (SNP microarrays)

[115-117]. Similar to CGH, both aCGH and SNP microarrays have the ability to detect changes in copy number over large regions of the genome. SNP microarrays further have the ability to genotype individuals at the probed sites, which may be useful in the detection of low-level somatic events [118]. Array-based approaches offer increased sensitivity over the entire genome for small CNVs relative to genome-wide microscopy-based approaches. aCGH and SNP microarray analysis can resolve regions less than 100 kb in size. However, the sensitivity of array-based approaches for somatic CNVs is dependent on having at least 5%–10% of the cells assayed containing the genetic variant. For larger CNVs affecting a smaller fraction of cells, microscopy-based approaches are more sensitive.

In both aCGH and SNP microarrays, deviations in relative probe intensities indicate deletion or insertion events. Normalized probe intensities are commonly reported as log-R ratios with higher intensities indicating insertions while lower intensities indicate deletions. For SNP microarrays, the relative intensities of the two probes (one specific to each allele) at a locus is informative and normalization of these intensities is measured as a B-allele frequency. For normal diploid tissues, B-allele frequencies approximate 0.0, 0.5, and 1.0 for AA, AB, and BB genotypes, respectively, while Log-R ratios approximate 0 across diploid regions.

The hybridization of genomic DNA to microarrays is inherently noisy and can be subject to large batch effects [119]. Furthermore, individual probes or even whole arrays may have errors caused by faulty manufacture. Together these artifacts make the detection of statistically significant mosaic CNVs difficult, but many software packages exist to aid in the detection of these events. Numerous tools use hidden Markov Models to integrate B-allele frequency and Log-R ratio information for the detection of mosaic events including

PennCNV-2, GPHMM and MixHMM [120-122]. gBPCR uses an approach similar to the Bayesian piecewise constant regression for the detection of mosaic abnormalities but has a long runtime per sample [123]. Our lab developed triPOD which uses multiple algorithms for the detection of mosaic events and is unique in that it utilizes parental genotypes allowing for more sensitive detection of haplotype-specific mosaic abnormalities [118].

2.2.4. Second-Generation Sequencing

Second-generation sequencing techniques have revolutionized human genetics in the last decade. Sequencing is performed either on single cells, a discrete number of cells, or bulk tissue. In the typical sequencing experiment, DNA is extracted from the input material and is fragmented, end-repaired, size-selected, and sequenced with the end result being strings of inferred nucleotides and their respective quality scores [124]. This information is used to align the sequencing reads to a reference genome. Differences between the aligned reads and the reference can be used to infer genetic variants including single-nucleotide variants or polymorphisms (SNVs or SNPs), insertions, deletions, translocations, and retrotransposition events. Furthermore, the total number of reads aligned to certain regions of the genome can be used to infer copy-number changes [125,126]. Numerous variations on this basic approach exist and here we will discuss the methods most applicable for the detection of mosaic events.

Somatic genetic variants have been discovered via whole-exome or whole-genome sequencing of bulk tissue from paired affected and unaffected portions of the body [61,65,66,105]. Whole-exome sequencing relies upon an oligonucleotide array-based capture of DNA fragments corresponding to exonic regions to reduce the sequence from noncoding regions of the genome [127-129]. At similar depth, exome and whole-genome sequencing are

considered to have similar sensitivity for most pathogenic SNVs and small insertions or deletions. Whole-exome sequencing is considered less sensitive for the identification of medium to large insertions or deletions or the detection of copy-number changes by analysis of read depth due to introduced biases. However, exome sequencing experiments are typically carried out at higher depth due to the lower cost of the method.

Numerous software packages allow for the identification of somatic variants from second-generation sequence data. Somatic variant callers typically evaluate next-generation sequence data from paired tumor/normal (or other affected/unaffected) samples. Examples include VarScan2 [126], SomaticSniper [130], JointSNVMix [131], Strelka [132], and MuTect [133]. After removal of low-quality reads, sequences are aligned to a reference genome to generate aligned binary sequence alignment/map (BAM) files [134]. At least three approaches have been employed for the detection of SNVs and small insertions or deletions. (1) Allele frequencies can be compared. For examples, VarScan2 performs pairwise comparisons of base calls and normalized sequence depth at each position, accounting for factors such base quality scores, coverage and variant allele frequencies. (2) Bayesian comparison of joint diploid genotype likelihood can be estimated for both samples. The SomaticSniper algorithm calculates the statistical significance of all somatic variants at positions above a minimum threshold of coverage using this method. (3) Other Bayesian approaches have been applied. For example, Strelka models the normal sample as germline variation plus noise, while the affected sample includes noise along with germline and somatic variation. Other types of somatic variation may be detected from bulk sequencing. Tools such as VarScan2, ADTeX, Control-FREEC, SomatiCA, and LUMPY may be used for the detection of somatic CNVs or structural variants [126,135-138].

Besides variant identification, quantification of the fraction of cells affected by particular somatic changes provides a better understanding of the extent of the mosaic mutation and the period during development at which it occurred. Several tools have been developed to deconvolute somatic mutations into distinct populations as reviewed by Yadav and De and Ding et al. [139,140].

An alternative approach to sequencing bulk tissue is sequencing single cells or small numbers of cells. Amplification can greatly increase the total amount of available DNA for sequencing at the expense of introduced biases such as allele dropout and chimeric amplification of genomic fragments [84,141-143]. Despite these introduced biases, amplification and subsequent second-generation sequencing or array-based analysis of single cells has been used to reliably find somatic copy number variation and retrotransposition events within the human brain as well as to map cell lineage within a bulk tumor dissection [58,59,84,144]. Numerous groups have also used single-cell techniques to discover SNVs or indels in single cells, however, allelic dropout and chimeric amplification are more problematic for these analyses as biases can be reduced for analysis of CNVs by increasing bin sizes but are more difficult to account for in analysis of SNVs [145-147].

2.3 Somatic Mosaicism in Disease

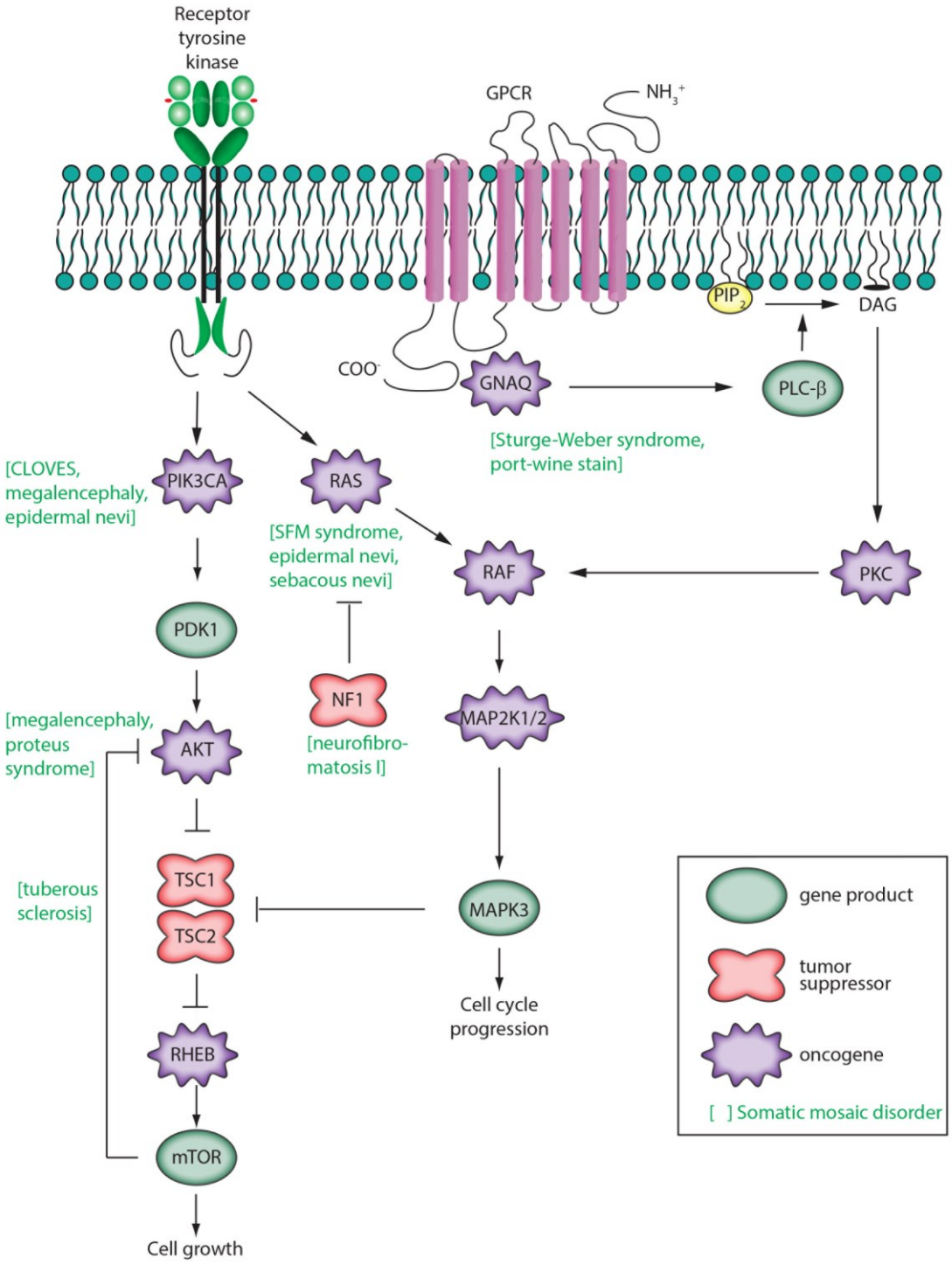
2.3.1. Cancer and Aging

The relationship between somatic mutation and cancer has been extensively reviewed elsewhere [12,20,97,148-150] and comprehensive lists of known oncogenes or tumor suppressors or genes significantly and recurrently mutated in cancer have been previously described [12,97]. Cancer has been described as having six hallmarks: proliferative signaling,

evading growth suppressors, resisting cell death, enabling replicative immortality, induction of angiogenesis, and inactivating invasion and metastasis [151]. Driver gene mutations are defined as conferring a selective growth advantage in tumor cells [12]. This may be achieved by elevating the activity of growth factors and/or their receptors, but more commonly driver mutations constitutively activate intracellular signal transduction cascades. Three of these are depicted in Figure 2.4 (in simplified form): Ras/Raf/MEK/ERK, Ras/PI3K/PTEN/Akt/mTOR [152], and *GNAQ*. These pathways contain both oncogenes (*RAS*, *RAF*, *MEK*, *PIK3CA*, *AKT*, *GNAQ*) and tumor suppressor genes (*NF1*, *PTEN*, *TSC1*, *TSC2*). For example the RAS family of oncogenes were the first oncogenes to be identified in cancer. Comprised of *HRAS*, *KRAS* and *NRAS*, activating mutations in these genes occur in approximately 20% of all cancers [153]. Germline variants are also well known to contribute to cancer morbidity [154-156]. Frequently, these variants affect proteins involved in DNA repair, highlighting the role of somatic mutations in tumorigenesis [157-160].

Figure 2.4. Signaling pathways in mosaic disease and cancer.

Three intracellular signaling pathways are shown schematically. **At left**, receptor tyrosine kinase activity leads to activation of PIK3CA, AKT, and mechanistic target of rapamycin (mTOR) [159,160]. mTOR participates in complexes (TORC1, activated by RHEB; TORC2, inhibited by RHEB) that regulate cell growth, proliferation, survival, and cell cycle progression. This pathway includes genes that are frequently mutated in tumors such as PIK3CA and PTEN (not shown). **At center**, secreted growth factors bind to receptor tyrosine kinase receptors on the cell surface leading to activation of the low molecular weight G protein Ras and subsequent activation of Raf, MEK 1/2, and ERK 1/2 (official gene symbols *MAPK3*, *MAPK1*). **At right**, a G-protein coupled receptor (GPCR) pathway is shown [157,158]. Ligands such as vasopressin, endothelin, glutamate, or norephinephrine bind to a GPCR. When bound by ligand, the receptor activates a G protein alpha subunit such as Gαq that binds and hydrolyzes GTP. This leads to activation of phospholipase Cβ producing inositol 1,4,5-triphosphate (IP_3) and membrane-associated diacylglycerol (DAG). DAG, through activation of protein kinase C, may activate the Raf/MEK/ERK pathway. IP_3 may bind to an IP_3 receptor activating calcium signaling pathways (not shown). Other G protein α subunits (such as Gαs encoded by *GNAS*) activate membrane-bound adenylate cyclase, producing cyclic AMP (cAMP) that activates protein kinase A (not shown).


Receptor tyrosine kinase
GPCR
NH_3^+
PIP_2
DAG
COO^-
GNAQ
PLC-β
[Sturge-Weber syndrome, port-wine stain]
[CLOVES, megalencephaly, epidermal nevi]
PIK3CA
RAS
[SFM syndrome, epidermal nevi, sebacous nevi]
RAF
PKC
PDK1
NF1
[neurofibro-matosis I]
MAP2K1/2
[megalencephaly, proteus syndrome]
AKT
[tuberous sclerosis]
TSC1
TSC2
MAPK3
Cell cycle progression
RHEB
mTOR
Cell growth
gene product
tumor suppressor
oncogene
[ ] Somatic mosaic disorder

In common solid tumors, ~95% of protein altering mutations consist of single base substitutions, >90% of which are missense mutations, <8% are nonsense mutations, and <2% affect splice sites or untranslated regions [12]. Relatively large numbers of somatic mutations occur in tumors that are associated with mutagens such as ultraviolet light and cigarette smoke. For example, in non-small cell lung carcinomas the average mutation frequency is greater than ten-fold higher in smokers compared to those who never smoke [161].

Large-scale projects and databases have been developed to provide comprehensive catalogues of somatic mutations found in cancer [162,163]. COSMIC (Catalogue Of Somatic Mutations In Cancer) includes information on more than 1.6 million mutations from nearly 1 million cancer samples and includes various types of mutations (fusions, genomic rearrangements, whole genomes, and copy number variants) [163].

The combination of well-characterized somatic mutation databases and low-cost sequencing technologies may lead to improved patient outcomes in the near future. Biopsied tumors may be screened rapidly for putative driver mutations based on cancer type, informing treatment. Furthermore, once a cancer is in remission, tumor-specific DNA may be assayed at low cost with ultra-sensitive second-generation sequencing-based techniques [164]. These advances will likely improve prognosis for millions of cancer patients within the next decade.

The primary risk factor for cancer is age, and cancers offer insight into age or mutagen-associated mutational processes [165]. Somatic mutations have long been suspected to be an important part of the molecular mechanism of aging, and accumulation of DNA lesions and mutations occurs in both the germline and soma over time [68,69,166,167]. By chance, these mutations may result in malignant transformation, apoptosis, or otherwise

hampered cellular function. As visible in cancers, the characteristics of acquired mutations differ by tissue type and are dependent upon environmental exposure [12]. Furthermore, frequently dividing stem cells and frequently transcribed genomic regions have different patterns of mutation that are cell-type specific.

In both mouse and human, increased rates of somatic mutation and numbers of DNA lesions due to either error-prone DNA polymerases or faulty DNA repair mechanisms cause cancer predisposition, early aging, and neurodegenerative phenotypes [20]. Increased rates of somatic mutation in the nuclear genome cause cancer predisposition, likely due to increased rates of mutation in somatic stem cell populations. This has been demonstrated in transgenic mice whose processive DNA polymerases lack proofreading. Notably, mice with mutated polymerases δ and ε develop distinct cancers but do not demonstrate premature aging phenotypes [168-170]. While these mice may not live long enough to demonstrate early aging phenotypes, their predisposition towards the development of cancer demonstrates a strong link between cancer and somatic mutation.

Mutations in genes affecting other pathways demonstrate a strong relationship between somatic mutations and aging. Mice with error-prone mitochondrial polymerases demonstrate a premature aging phenotype without cancer predisposition, although subsequent data by some of the same authors demonstrate that mitochondrial point mutations are unlikely the primary cause of aging in normal mice [171,172]. Individuals with defects in DNA repair also demonstrate symptoms of progeria. Cockayne syndrome (OMIM #216400) is caused by defects in transcription-coupled exonucleotide repair leading to an early aging phenotype combined with intellectual disability and neurodegeneration without noted predisposition to development of cancer [173]. Mutations in the genes encoding RecQ

helicases cause Werner syndrome (OMIM #277700) and Rothmund-Thomson syndrome (OMIM #268400) [174]. The most prominent phenotype of individuals affected by these diseases is premature aging, although these individuals are also predisposed to developing cancer [174]. Bloom Syndrome (OMIM #210900) is notable in that it is also caused by mutations in a RecQ helicase-like protein and also increases cancer incidence, but does not appear to result in progeria. Mutations in numerous other genes are known to cause cancer predisposition. One such example is *BUB1B*. Loss of BUB1B protein function leads to premature chromatid separation and mosaic variegated aneuploidy syndrome 1 (OMIM #257300) typically resulting in cancer predisposition and intellectual disability [175].

Cancer is associated with many genomic changes. Large chromosomal changes occur in a variety of noncancerous conditions. An example is Pallister-Killian syndrome (OMIM #601803) is a dysmorphic condition caused by mosaicism for tetrasomy 12p. Affected individuals display tissue mosaicism, typically with apparently normal karyotypes from lymphocytes but 47 chromosomes in skin fibroblasts and chorionic villus and amniotic fluid cells. The extra chromosome is an isochromosome for a portion of chromosome 12p. In several cases hexasomy of chromosome 12p has been observed.

2.3.2. Neurodegenerative Disease

Somatic mutation is suspected to have a role in neurodegenerative disease [20,21]. As in cancer, mutations in genes directly involved in DNA repair are implicated in neurodegenerative diseases such as ataxia-telangiectasia (OMIM #208900) and ataxia-ocular apraxia 1 (OMIM #208920) [19,174,176-179]. These neurodegenerative phenotypes are likely caused by an increase of somatic mutation in the nervous system leading to cellular dysfunction and indicate a possible role for somatic changes and DNA lesions in age-related

related neurodegenerative disorders.

There is evidence that mosaic mutations or accumulated damage to other macromolecules play a role in Alzheimer's disease (OMIM #104300) and Creutzfeldt-Jakob disease (CJD) (OMIM #123400). Alzheimer's disease is characterized by the accumulation of β-amyloid (Aβ) plaques while CJD is caused by misfolded protein PRNP [180,181]. Significant incidence of both diseases is attributed to familial risk and causal mosaic mutations have been found in sporadic cases [182,183]. Aβ plaques have long been implicated in the formation of prions and introduction of Aβ plaques into the brains of mice overexpressing Aβ leads to disease progression [184-186]. Consistent with the link to prions, the pathology of inoculated mice displays phenotypes dependent upon the infecting host [186]. This has been corroborated by more recent experiments, which demonstrate that Aβ aggregates from distinct sources have unique biophysical characteristics depending on the seeding protein [187-189]. While it is possible that sporadic misfolded or damaged proteins act as seeds in Alzheimer's, this is unlikely given the steep increase in disease incidence later in life and the constant turnover of cellular proteins [190]. This steep rise in incidence mirrors the rise in incidence of CJD in individuals who have predisposing mutations [191]. It is possible that in both diseases misfolded proteins arising as a result of age-related somatic mutation or damage to other macromolecules in single cells act as seeds for the initial protein aggregates.

2.3.3. Monogenic Disease

A list of diseases suspected to be caused by obligatory somatic mutations has been previously described [22] and subsequently updated [23,24]. We note that somatic mutation likely contributes significantly to nearly all Mendelian diseases.

We have described a series of oncogenes and tumor suppressor genes that undergo somatic mutation in cancer. These same genes can also acquire somatic mutations that result in neurocutaneous disorders or overgrowth syndromes, depending the particular cell type and developmental stage at which the mutation occurs. Mutations in *GNAQ* cause Sturge-Weber syndrome and port-wine stain birthmarks as well as uveal melanoma, as discussed above. Similarly, somatic mutations in *GNAS* can cause McCune-Albright syndrome or benign tumors such as adenomas. We next highlight several specific examples of such disorders affecting genes encoding intracellular signaling pathways (Figure 2.4).

Phosphatidylinositol 3-kinases (PIK3s) are lipid kinases that phosphorylate phosphatidylinositol and other phosphoinositides, catalyzing intracellular signaling pathways involving a PI3K/AKT/mTOR network (Figure 2.4). Somatic, mosaic, gain-of-function mutations in *PIK3CA* (OMIM *171834) are associated with several syndromes involving overgrowth of the brain or lipomatous body overgrowth [192]. These include CLOVE (an acronym for congenital lipomatous overgrowth, vascular anomalies, and epidermal nevi) syndrome, megalencephaly-capillary malformation syndrome, fibroadipose hyperplasia, and hemimegalencephaly. These conditions are often characterized by early segmental overgrowth, abnormal vasculogenesis, digital anomalies, cortical brain malformations, and connective tissue dysplasia. Somatic gain-of-function mutations in *PIK3CA* are also found in a broad range of cancers (ovarian, breast, lung, stomach, colorectal, and brain). While over 100 activating mutations in *PIK3CA* are known, mutations in two domains of the protein account for 80% of cancer-associated somatic mutations, and these same sites can be mutated in overgrowth disorders [193].

Clinical presentation of Proteus syndrome (OMIM #176920) includes distorting, progressive overgrowth of various tissues including skin, skeleton, adipose, and central nervous system. In most patients it is caused by somatic mosaic mutation of *AKT1* involving c.49G>A (p.Glu17Lys) [66]. This identical mutation was previously known to be associated with breast, colorectal and ovarian cancers [194]. The holomlogues of *AKT1*; *AKT2* and *AKT3* are also known to cause somatic disorder. p.Glu17Lys mutations in *AKT3* cause hemimegalencephaly and other brain malformations, while the identical mutation in *AKT2* is causative for hypoglycemia [65,67,195,196].

Germline inactivating mutations in the *TSC1* gene encoding hamartin cause tuberous sclerosis 1 (OMIM #191100), while mutations in *TSC2* encoding tuberin cause tuberous sclerosis 2 (OMIM #613254). Hamartin and tuberin to act as tumor suppressors by activating the GTPase function of RHEB [197]. Inactivating mutations in a single allele are sufficient to cause tuberous sclerosis as rare somatic inactivating mutations, lack of expression of the second allele or mosaic UDP events give rise to the multiple benign tumors, tubers and macules characteristic of the disease [198,199].

Neurofibromatosis 1 (OMIM #162200) (NF1) is characterized by the occurrence of at least two (of a list of seven) features such as café au lait spots, cutaneous neurofibromas, Lisch nodules (hamartomas) of the iris, and inguinal freckles [200]. Clinical diagnosis requires a first-degree relative with the condition. It is inherited in an autosomal dominant manner (and is among the most common such disorders with a prevalence of 1:3000). Most cases of NF1 are caused by heterozygous loss-of-function mutations of the tumor suppressor gene encoding neurofibromin 1. But only 50% of NF1 individuals have an affected parent, with another 50% having a de novo mutation. Neurofibromin 1 is a negative regulator of the

ras signal transduction pathway, with loss of function mutations in neurofibromin 1 leading to RAS activation.

It is possible that mosaic variation occurring during development may result in disease across numerous tissues. One such example is somatic mutation of *IDH1* and *IDH2* that has been shown to cause Ollier disease and Maffucci syndrome. These syndromes are characterized by multiple enchondromas (benign bone tumors originating from cartilage). The causative variants for disease are typically not detectable outside of the tumors indicating that relatively few cells harbor the mutation [201].

The application of sensitive approaches for the detection of mosaicism to a smaller subset of genes based on a patient's phenotype may increase the likelihood of finding causative variants. Jamuar *et al.* applied this approach examining two sets of previously implicated genes in 158 individuals with cerebral cortical defects. Causal mutations were found in 27 individuals, eight of who harbored the causative variant in a mosaic fashion. Notably, causal mutations were only validated at extremely high read depth (>500×) highlighting both the importance of sequence coverage for the detection of mosaic variation and the utility of targeted approaches [202].

Somatic mutations are also known to cause reversion to normal mutations in individuals with monogenic disease [25,26,28,203,204]. Revertant mosaicism occurs when cells harboring a disease-causing mutation revert *in vivo* to a wild-type allele. The disease-causing mutation could be inherited from the germline or somatic. This has been observed for heritable skin diseases such as ichthyosis with confetti (OMIM #609165) and epidermolysis bullosa (OMIM #226650) [203,205] as well as rare blood disorders such as Fanconi anemia (OMIM #227646) and severe combined immunodeficiency resulting from adenosine

deaminase deficiency (OMIM #102700) [206,207]. These somatic reversions to normal events may significantly ameliorate disease symptoms if the reversion occurs early enough in development.

For many other overgrowth syndromes somatic mutations have yet to be identified. Examples include Klippel-Trenaunay-Weber syndrome (OMIM %149,000), which involves cutaneous hemangiomata and clinically resembles Sturge-Weber syndrome; and Cobb syndrome (cutaneomeningospinal angiomatosis), which involves vascular cutaneous, muscular, osseous, or other lesions of spinal segments.

2.3.4. Complex Disease

Multiple recent papers have proposed that somatic mutation may play a role in the etiology of complex disease [6,208,209]. Studies of simplex autism probands have determined that de novo mutations account for 10%–20% of disease incidence and that at least 30% of de novo mutations can be causally implicated in simplex cases [210-213]. With de novo mutations playing such a large role, it is likely that post-zygotic somatic variation also contributes to disease in some individuals. To date, most genetic analysis has found few genetic variants to explain complex disease incidence, suggesting the occurrence of "missing heritability" [214]. A possible model is that somatic variation occurs in conjunction with common and rare inherited variation to cause disease. While this model is not directly supported by current evidence, recent experiments indicate that it warrants investigation. One surprising result from in situ hybridization experiments on postmortem brain tissue is the increased presence of patches of cortical disorganization in individuals with autism relative to controls [215]. The authors note that they examined only a small subsection of the brain and therefore cortical disorganization is likely widespread in individuals with autism.

Furthermore, an interesting conclusion of recent large-scale examination of exonic de novo mutations in simplex autism is that most de novo variation implicated as causal occurs opposite wild type alleles [212]. Given that large CNVs are common in neurons of the cortex [83,85], we propose a model of brain-specific somatic mutation occurring opposite inherited de novo or rare mutation resulting in sporadic brain-specific loss of gene function and patches of cortical disorganization.

2.4. Conclusions

While the role of somatic mosaicism in disease currently under active investigation, it is clear that functional somatic mosaicism has a significant role in human disease. In the last decade, major advances in both cytogenetic and second-generation sequencing techniques have enabled researchers to discover causative somatic mutations for an increasing number of diseases, and driver mutations in an increasing number of cancers. Furthermore, this increased understanding of the genetic underpinnings of disease is likely to lead to improved patient outcomes in the near future.

Chapter 3: The Contribution of Mosaic Variants to Autism Spectrum Disorder

3.1 Introduction

DNA is constantly exposed to natural and artificial mutagenic processes and therefore continually develops lesions and undergoes subsequent error-prone repair. In multi-cellular organisms, these mutations may arise at any time during development resulting in diverse organismal and cellular phenotypes, including disease. The severity of these phenotypes is dependent upon not only the particular genetic change but also the affected cell type and time in development at which the mutation occurs. Obligatory somatic disorders, in which prenatally lethal germline mutations occur post-zygotically, are one extreme [29].

In contrast to obligatory somatic mutation, *de novo* mutation is thought to primarily occur in the parental germline, typically resulting in genetic variation that is heterozygous in every cell of an organism. Such mutation is *de novo* in the sense that it is below the limit of detection in a parental sample (usually DNA derived from blood). An early report using comparative genomic hybridization indicated that large *de novo* copy-number variants are enriched in ASD probands [216]. From these results it was hypothesized, and subsequent microarray and whole-exome sequencing experiments have shown, that a substantial fraction of genetic liability arises *de novo* in every generation [212,217-228].

The exact developmental time at which *de novo* mutations occur however, is under active investigation. Some *de novo* variants discovered though whole-exome sequencing have properties consistent with mosaicism [222,225,229,230]. Recent experiments using high-depth targeted sequencing have indicated that eight of 27 likely causal variants in individuals with cortical malformations are present as mosaics, occasionally at very low

alternate-allele reads fractions (AARF) [202]. Mosaic mutations have been found to occur in single individuals of monozygotic twin pairs [231,232]. Furthermore, 6.5% of identified *de novo* mutations in individuals with severe intellectual disability occur as mosaics [233]. Here we show that *de novo* variation in a large whole-exome sequencing dataset is frequently mosaic and that such mosaic variation is likely to contribute to disease diagnoses in some affected individuals.

3.2 Materials and Methods

3.2.1 Paired sample whole-exome sequencing

Paired samples were obtained from the University of Maryland Brain and Tissue Bank as detailed in **Table 3.1**. Individuals were diagnosed with ASD (n=12) or were controls; criteria for diagnosing ASD included the Autism Diagnostic Interview-Revised (ADI-R), Childhood Autism Rating Scale (CARS), and Autism Diagnostic Observation Schedule (ADOS) as detailed **Table 3.1**. DNA was extracted from tissue dissections according to protocols in the QIAGEN Genomic DNA Handbook. Exonic regions were selectively captured using Agilent SureSelectXT Human All Exon V5. Sequencing was performed at the Center for Inherited Disease Research at Johns Hopkins generating 100 bp sequence reads on an Illumina HiSeq. CIDRSeqSuite version 3.0.1 was used for processing of the raw data files. BCL files were converted to qseq format using Illumina's BCL converter. qseq files were then demultiplexed and converted to FASTQ files using a custom demultiplexer. Paired-end alignment was performed using BWA aln to the 1000 genomes hg19/GRCh37 reference genome [234]. SAM files were sorted, converted to BAM, and duplicates were marked with Picard. GATK was used for local realignment and base quality score recalibration [235,236]. Quality metrics for these data are provided in **Table 3.2**.

Table 3.1 Tissue samples obtained from the University of Maryland Brain and Tissues Bank.

PMI, postmortem interval.

Brain Bank ID	Diagnosis	Age		Gender	Ethnicity	PMI	Storage		ADI-R Sect.A	Sect.B verbal	Sect.B nonverbal	Sect.C	Sect.D
		Years	Days			Hours	Years	Days					
1349	ASD	5	220	M	Caucasian	39	13	268					
4722	Control	14	198	M	Caucasian	16	9	234	N/A	N/A	N/A	N/A	N/A
4671	ASD	4	165	F	African American	13	9	226	26	n/a	13	3	5
4849	ASD	7	171	M	African American	20	8	243	22	18	n/a	8	3
4907	Control	4	274	F	African American	15	7	69	N/A	N/A	N/A	N/A	N/A
4999	ASD	20	274	M	Caucasian	14	6	22	n/a	n/a	n/a	n/a	n/a
5144	ASD	7	55	M	Caucasian	3	6	316	28	20	12	3	3
5278	ASD	15	324	F	Caucasian	13	5	310	22	21	11	5	5
5391	Control	8	286	M	Caucasian	12	4	258	N/A	N/A	N/A	N/A	N/A
5408	Control	6	309	M	African American	16	4	145	N/A	N/A	N/A	N/A	N/A
5419	ASD	19	350	F	Caucasian	22	4	167	24	14	n/a	6	3
797	ASD	9	100	M	Caucasian	13	17	178	24	20	13	6	blank
4899	ASD	14	126	M	Caucasian	9	8	183	22	n/a	14	8	4
5176	ASD	22	199	M	African American	18	6	224	25	13	13	7	5
5115	ASD	46	135	M	Caucasian	29	6	358	27	n/a	14	8	1
5403	ASD	16	266	M	Caucasian	35	4	235	30	n/a	14	5	5

Table 3.2. Quality control metrics of whole-exome sequence data from paired samples as reported by CIDR.

TARGET TERRITORY is the size of the exome-capture target. PCT PF UQ READS ALIGNED is the number of unique reads passing vendor quality filters that were successfully aligned to the genome. MEAN TARGET COVERAGE is the mean coverage over exome-capture targets. ZERO CVG TARGETS PCT is the fraction of targets without coverage over any base.

IDENTIFIER	TARGET TERRITORY	TOTAL READS	PCT PF UQ READS ALIGNED	MEAN TARGET COVERAGE	ZERO CVG TARGETS PCT	PCT TARGET BASES 10X	PCT TARGET BASES 20X	PCT TARGET BASES 30X	MEAN INSERT SIZE
1349_brain	36796199	72883652	0.976	71.6	0.0061	0.97	0.92	0.83	290
1349_kidney	36796199	104449644	0.966	97.6	0.0053	0.98	0.96	0.91	291
4722_brain	36796199	115826574	0.973	107.3	0.0052	0.98	0.96	0.93	294
4722_heart	36796199	80539102	0.973	76.8	0.0063	0.97	0.93	0.86	299
4671_brain	36796199	100995106	0.971	91.9	0.0071	0.97	0.95	0.89	304
4671_heart	36796199	116486392	0.973	141.4	0.0077	0.98	0.95	0.92	208
4849_brain	36796199	82153876	0.956	72.2	0.0055	0.98	0.92	0.84	307
4849_heart	36796199	80847910	0.971	70.7	0.0054	0.97	0.92	0.84	300
4907_brain	36796199	76114756	0.974	71.7	0.0067	0.97	0.92	0.83	299
4907_heart	36796199	102900278	0.973	100.1	0.0062	0.98	0.95	0.91	289
4999_brain	36796199	77877604	0.976	75.2	0.0056	0.98	0.93	0.85	307
4999_heart	36796199	103195310	0.976	100.8	0.0052	0.98	0.96	0.92	296
5144_brain	36796199	93790024	0.969	88.5	0.0058	0.98	0.94	0.89	295
5144_heart	36796199	85663888	0.973	78.7	0.0052	0.98	0.94	0.87	308
5278_brain	36796199	97717988	0.974	95.7	0.0066	0.98	0.95	0.91	288
5278_kidney	36796199	89999006	0.973	82.5	0.0063	0.98	0.94	0.88	306
5391_brain	36796199	90248082	0.971	81.5	0.0051	0.98	0.94	0.88	312

IDENTIFIER	TARGET TERRITORY	TOTAL READS	PCT PF UQ READS ALIGNED	MEAN TARGET COVERAGE	ZERO CVG TARGETS PCT	PCT TARGET BASES 10X	PCT TARGET BASES 20X	PCT TARGET BASES 30X	MEAN INSERT SIZE
5391_heart	36796199	84728178	0.973	82.2	0.0053	0.98	0.94	0.87	290
5408_brain	36796199	97274040	0.972	91.9	0.0052	0.98	0.95	0.90	297
5408_heart	36796199	99165538	0.955	88.3	0.0054	0.98	0.95	0.89	298
5419_brain	36796199	106813158	0.976	101.1	0.0067	0.98	0.96	0.92	300
5419_heart	36796199	99144944	0.974	89.7	0.0065	0.98	0.95	0.90	309
4889_brain	36796199	124350560	0.971	161.0	0.0078	0.97	0.94	0.90	169
4889_heart	36796199	106054250	0.975	109.0	0.0053	0.98	0.96	0.92	253
5115_brain	36796199	117191886	0.968	106.4	0.0048	0.98	0.96	0.93	301
5115_heart	36796199	58995338	0.977	58.6	0.0061	0.97	0.89	0.77	283
5179_brain	36796199	55503602	0.970	52.9	0.0062	0.96	0.87	0.74	296
5179_heart	36796199	78292648	0.975	100.4	0.0076	0.97	0.92	0.87	180
5403_brain	36796199	101087994	0.974	95.6	0.0050	0.98	0.96	0.91	287
5403_heart	36796199	102836624	0.974	97.0	0.0050	0.98	0.96	0.91	279
797_brain	36796199	149786604	0.971	173.2	0.0063	0.98	0.97	0.95	207
797_heart	36796199	104011488	0.971	115.2	0.0060	0.98	0.96	0.92	225

3.2.2 Paired sample tissue-specific variant calling

Tissue-specific variants were called from paired samples using MuTect 2.7-1 for SNV discovery and Strelka 1.0.13 for indel discovery [132,133]. Input to these programs requires specifying a "tumor" and a "normal" sample. For each paired sample, variants were called twice so that mosaic variants were identified in both the brain and heart/kidney tissue. Validation of these variant calls was performed as indicated in **Table 3.3** (run 1), with variants with the most severe functional effect selected for validation.

Table 3.3. Amplicon-targeted sequencing of potential tissue-specific variants in paired samples.

Sample	Run Number	Chromosome	Position	Ref	Alt	Total_depth	Variant_reads	Percent_variants
1349B	1	3	123376027	C	T	8125	36	0.004430769
1349K	1	3	123376027	C	T	8098	36	0.004445542
4671B	1	3	123376027	C	T	8121	51	0.006280015
4671H	1	3	123376027	C	T	8116	49	0.006037457
4671K	1	3	123376027	C	T	8115	31	0.003820086
4722B	1	3	123376027	C	T	8117	50	0.006159911
4722H	1	3	123376027	C	T	8120	38	0.004679803
4722K	1	3	123376027	C	T	1	0	0
4849B	1	3	123376027	C	T	8121	28	0.003447851
4849H	1	3	123376027	C	T	8125	44	0.005415385
4899B	1	3	123376027	C	T	8119	52	0.00640473
4899H	1	3	123376027	C	T	8113	41	0.005053618
4907B	1	3	123376027	C	T	8116	39	0.004805323
4907H	1	3	123376027	C	T	8121	39	0.004802364
4907K	1	3	123376027	C	T	8119	38	0.004680379
4999B	1	3	123376027	C	T	8120	45	0.005541872
4999H	1	3	123376027	C	T	8110	45	0.005548705
5115B	1	3	123376027	C	T	8099	38	0.004691937
5115K	1	3	123376027	C	T	0	0	0
5144B	1	3	123376027	C	T	8113	33	0.004067546
5144H	1	3	123376027	C	T	8120	49	0.006034483
5155H	1	3	123376027	C	T	8129	27	0.003321442
5176B	1	3	123376027	C	T	8111	43	0.005301442
5176H	1	3	123376027	C	T	8099	39	0.004815409
5278B	1	3	123376027	C	T	8113	58	0.00714902

Sample	Run Number	Chromosome	Position	Ref	Alt	Total_depth	Variant_reads	Percent_variants
5278K	1	3	123376027	C	T	8111	46	0.005671311
5391B	1	3	123376027	C	T	8110	32	0.003945746
5391H	1	3	123376027	C	T	8111	35	0.004315128
5403B	1	3	123376027	C	T	8110	19	0.002342787
5408B	1	3	123376027	C	T	8119	44	0.005419387
5408H	1	3	123376027	C	T	0	0	0
5419B	1	3	123376027	C	T	8119	36	0.004434044
5419H	1	3	123376027	C	T	8120	36	0.004433498
797B	1	3	123376027	C	T	8120	41	0.005049261
797H	1	3	123376027	C	T	8114	37	0.00456002
1349B	1	10	79397364	C	T	0	0	0
1349K	1	10	79397364	C	T	0	0	0
4671B	1	10	79397364	C	T	0	0	0
4671H	1	10	79397364	C	T	0	0	0
4671K	1	10	79397364	C	T	0	0	0
4722B	1	10	79397364	C	T	0	0	0
4722H	1	10	79397364	C	T	0	0	0
4722K	1	10	79397364	C	T	0	0	0
4849B	1	10	79397364	C	T	0	0	0
4849H	1	10	79397364	C	T	0	0	0
4899B	1	10	79397364	C	T	0	0	0
4899H	1	10	79397364	C	T	0	0	0
4907B	1	10	79397364	C	T	0	0	0
4907H	1	10	79397364	C	T	0	0	0
4907K	1	10	79397364	C	T	0	0	0
4999B	1	10	79397364	C	T	0	0	0
4999H	1	10	79397364	C	T	0	0	0

Sample	Run Number	Chromosome	Position	Ref	Alt	Total_depth	Variant_reads	Percent_variants
5115B	1	10	79397364	C	T	0	0	0
5115K	1	10	79397364	C	T	0	0	0
5144B	1	10	79397364	C	T	0	0	0
5144H	1	10	79397364	C	T	0	0	0
5155H	1	10	79397364	C	T	0	0	0
5176B	1	10	79397364	C	T	0	0	0
5176H	1	10	79397364	C	T	0	0	0
5278B	1	10	79397364	C	T	0	0	0
5278K	1	10	79397364	C	T	0	0	0
5391B	1	10	79397364	C	T	0	0	0
5391H	1	10	79397364	C	T	0	0	0
5403B	1	10	79397364	C	T	0	0	0
5408B	1	10	79397364	C	T	0	0	0
5408H	1	10	79397364	C	T	0	0	0
5419B	1	10	79397364	C	T	0	0	0
5419H	1	10	79397364	C	T	0	0	0
797B	1	10	79397364	C	T	0	0	0
797H	1	10	79397364	C	T	0	0	0
1349B	1	15	28370205	T	G	6648	139	0.020908544
1349K	1	15	28370205	T	G	6541	630	0.096315548
4671B	1	15	28370205	T	G	6545	151	0.023071047
4671H	1	15	28370205	T	G	6852	303	0.044220665
4671K	1	15	28370205	T	G	6757	255	0.037738641
4722B	1	15	28370205	T	G	6773	376	0.055514543
4722H	1	15	28370205	T	G	6663	155	0.023262795
4722K	1	15	28370205	T	G	0	0	0
4849B	1	15	28370205	T	G	6722	138	0.020529604

Sample	Run Number	Chromosome	Position	Ref	Alt	Total_depth	Variant_reads	Percent_variants
4849H	1	15	28370205	T	G	6734	149	0.022126522
4899B	1	15	28370205	T	G	6609	294	0.044484793
4899H	1	15	28370205	T	G	6983	627	0.089789489
4907B	1	15	28370205	T	G	6781	372	0.054859165
4907H	1	15	28370205	T	G	6577	238	0.036186711
4907K	1	15	28370205	T	G	6606	245	0.037087496
4999B	1	15	28370205	T	G	6624	256	0.038647343
4999H	1	15	28370205	T	G	7018	679	0.096751211
5115B	1	15	28370205	T	G	6898	956	0.138590896
5115K	1	15	28370205	T	G	0	0	0
5144B	1	15	28370205	T	G	6837	482	0.070498757
5144H	1	15	28370205	T	G	6819	381	0.055873295
5155H	1	15	28370205	T	G	6747	158	0.023417815
5176B	1	15	28370205	T	G	6785	519	0.076492262
5176H	1	15	28370205	T	G	7062	703	0.099546871
5278B	1	15	28370205	T	G	6970	478	0.068579627
5278K	1	15	28370205	T	G	6849	486	0.070959264
5391B	1	15	28370205	T	G	6913	684	0.098944019
5391H	1	15	28370205	T	G	7020	646	0.092022792
5403B	1	15	28370205	T	G	6894	504	0.07310705
5408B	1	15	28370205	T	G	6853	409	0.059681891
5408H	1	15	28370205	T	G	0	0	0
5419B	1	15	28370205	T	G	6668	236	0.035392921
5419H	1	15	28370205	T	G	6645	221	0.033258089
797B	1	15	28370205	T	G	6628	240	0.036210018
797H	1	15	28370205	T	G	6875	443	0.064436364
1349B	1	19	12772090	C	T	41	0	0

Sample	Run Number	Chromosome	Position	Ref	Alt	Total_depth	Variant_reads	Percent_variants
1349K	1	19	12772090	C	T	11	1	0.090909091
4671B	1	19	12772090	C	T	12	11	0.916666667
4671H	1	19	12772090	C	T	10	10	1
4671K	1	19	12772090	C	T	29	25	0.862068966
4722B	1	19	12772090	C	T	25	0	0
4722H	1	19	12772090	C	T	9	1	0.111111111
4722K	1	19	12772090	C	T	0	0	0
4849B	1	19	12772090	C	T	43	19	0.441860465
4849H	1	19	12772090	C	T	16	12	0.75
4899B	1	19	12772090	C	T	11	7	0.636363636
4899H	1	19	12772090	C	T	13	6	0.461538462
4907B	1	19	12772090	C	T	16	9	0.5625
4907H	1	19	12772090	C	T	19	9	0.473684211
4907K	1	19	12772090	C	T	8	6	0.75
4999B	1	19	12772090	C	T	13	8	0.615384615
4999H	1	19	12772090	C	T	17	10	0.588235294
5115B	1	19	12772090	C	T	17	0	0
5115K	1	19	12772090	C	T	0	0	0
5144B	1	19	12772090	C	T	15	15	1
5144H	1	19	12772090	C	T	8	8	1
5155H	1	19	12772090	C	T	31	1	0.032258065
5176B	1	19	12772090	C	T	8	0	0
5176H	1	19	12772090	C	T	12	1	0.083333333
5278B	1	19	12772090	C	T	3	3	1
5278K	1	19	12772090	C	T	5	5	1
5391B	1	19	12772090	C	T	4	1	0.25
5391H	1	19	12772090	C	T	6	0	0

Sample	Run Number	Chromosome	Position	Ref	Alt	Total_depth	Variant_reads	Percent_variants
5403B	1	19	12772090	C	T	10	0	0
5408B	1	19	12772090	C	T	5	0	0
5408H	1	19	12772090	C	T	0	0	0
5419B	1	19	12772090	C	T	13	0	0
5419H	1	19	12772090	C	T	24	0	0
797B	1	19	12772090	C	T	11	10	0.909090909
797H	1	19	12772090	C	T	20	20	1
1349B	1	19	49926533	G	C	8001	217	0.02712161
1349K	1	19	49926533	G	C	1998	67	0.033533534
4671B	1	19	49926533	G	C	3768	129	0.034235669
4671H	1	19	49926533	G	C	2056	67	0.032587549
4671K	1	19	49926533	G	C	3553	123	0.034618632
4722B	1	19	49926533	G	C	4066	139	0.034185932
4722H	1	19	49926533	G	C	3600	132	0.036666667
4722K	1	19	49926533	G	C	0	0	0
4849B	1	19	49926533	G	C	4255	139	0.03266745
4849H	1	19	49926533	G	C	3514	124	0.035287422
4899B	1	19	49926533	G	C	2611	78	0.029873612
4899H	1	19	49926533	G	C	1558	41	0.026315789
4907B	1	19	49926533	G	C	4218	114	0.027027027
4907H	1	19	49926533	G	C	3163	89	0.028137844
4907K	1	19	49926533	G	C	3432	87	0.02534965
4999B	1	19	49926533	G	C	6129	187	0.030510687
4999H	1	19	49926533	G	C	1847	49	0.026529507
5115B	1	19	49926533	G	C	2628	86	0.032724505
5115K	1	19	49926533	G	C	0	0	0
5144B	1	19	49926533	G	C	3039	75	0.024679171

Sample	Run Number	Chromosome	Position	Ref	Alt	Total_depth	Variant_reads	Percent_variants
5144H	1	19	49926533	G	C	2055	55	0.02676399
5155H	1	19	49926533	G	C	7090	193	0.027221439
5176B	1	19	49926533	G	C	2277	82	0.036012297
5176H	1	19	49926533	G	C	1913	42	0.021955044
5278B	1	19	49926533	G	C	3657	117	0.031993437
5278K	1	19	49926533	G	C	3779	112	0.02963747
5391B	1	19	49926533	G	C	3078	74	0.024041585
5391H	1	19	49926533	G	C	2200	69	0.031363636
5403B	1	19	49926533	G	C	2898	78	0.026915114
5408B	1	19	49926533	G	C	4385	104	0.023717218
5408H	1	19	49926533	G	C	0	0	0
5419B	1	19	49926533	G	C	4108	120	0.029211295
5419H	1	19	49926533	G	C	4319	133	0.030794165
797B	1	19	49926533	G	C	2863	88	0.030736989
797H	1	19	49926533	G	C	1874	61	0.032550694
1349B	2	9	131020812	C	A	1993	52	2.542787286
4907K	2	9	131020812	C	A	328	282	46.2295082
5408K	2	9	131020812	C	A	658	628	48.83359253
4907H	2	9	131020812	C	A	1860	15	0.8
5419H	2	4	93225845	G	T	52469	330	0.625011837
4671H	2	4	93225845	G	T	42184	182	0.429589765
5419B	2	4	93225845	G	T	53867	299	0.552006794
4999H	2	13	113742714	G	A	55469	197	0.353896454
1349B	2	13	113742714	G	A	373	5	1.322751323
5419H	2	11	47600657	C	T	43378	192	0.440670186
5176K	2	11	47600657	C	T	44861	145	0.322179265
5419K	2	11	47600657	C	T	33171	101	0.303558548

Sample	Run Number	Chromosome	Position	Ref	Alt	Total_depth	Variant_reads	Percent_variants
5176B	2	11	47600657	C	T	24628	73	0.295534594
4671H	2	11	47600657	C	T	25866	73	0.281429508
5408K	2	11	47600657	C	T	28463	108	0.37800567
4899H	2	11	47600657	C	T	31216	106	0.33842028
5419B	2	11	47600657	C	T	26031	113	0.432221542
5176H	2	11	47600657	C	T	35327	118	0.332910142
4722H	2	7	56149352	G	A	23996	28	0.116550117
5419K	2	7	56149352	G	A	13492	27	0.199718914
797H	2	1	151238513	C	T	58590	118	0.200994754
4849H	2	1	151238513	C	T	32627	76	0.232394582
4849K	2	1	151238513	C	T	23183	67	0.288172043
5722H	2	17	74271953	C	T	22717	116	0.508036614
797K	2	17	74271953	C	T	4529	35	0.766871166
4722K	2	17	74271953	C	T	16932	80	0.47025629
5176K	2	16	67289691	G	A	47167	108	0.228450555
5176B	2	16	67289691	G	A	40377	93	0.229799852
5115H	2	12	123950169	G	A	61580	196	0.317275317
4671K	2	12	123950169	G	A	42831	104	0.242226622
4671B	2	12	123950169	G	A	31102	315	1.002641882
4671B	2	12	123950169	G	A	69022	151	0.218293265
4671H	2	12	123950169	G	A	52956	249	0.468001128
4671K	2	12	123950169	G	A	39577	69	0.174040256
5176K	2	22	39811081	G	A	332	1	0.3003003
5419H	2	22	39811081	G	A	70283	197	0.279511918
5391K	2	22	39811081	G	A	52660	97	0.183861857
5391K	2	22	39811081	G	A	52916	82	0.15472282
5419B	2	16	426147	G	A	26280	38	0.144387871

Sample	Run Number	Chromosome	Position	Ref	Alt	Total_depth	Variant_reads	Percent_variants
5419H	2	16	426147	G	A	391	1	0.255102041
5144H	2	16	426147	G	A	38942	80	0.205012557
5419K	2	16	426147	G	A	52727	64	0.121232786
797K	2	7	116955170	ATCCTT	A	34412	279	0.804243175
5403K	2	7	116955170	ATCCTT	A	28513	207	0.720752089
5115K	2	7	116955170	ATCCTT	A	30198	213	0.700404459
5144B	2	7	116955170	ATCCTT	A	45784	395	0.855367158
5144H	2	7	116955170	ATCCTT	A	516	3	0.578034682
5278B	2	7	116955170	ATCCTT	A	56898	416	0.725826151
5408K	2	7	116955170	ATCCTT	A	30900	198	0.636696894
5278K	2	7	116955170	ATCCTT	A	24550	159	0.643490226
4899B	2	7	116955170	ATCCTT	A	7426	50	0.668806849
4899H	2	7	116955170	ATCCTT	A	49376	362	0.727813744
5176H	2	7	116955170	ATCCTT	A	44270	289	0.648578289
4899K	2	7	116955170	ATCCTT	A	36772	232	0.626959248
797B	2	7	116955170	ATCCTT	A	37251	228	0.608340671
5403B	2	7	116955170	ATCCTT	A	41911	289	0.684834123

To more carefully examine the properties of the mosaic variants called by MuTect, variants were recalled jointly in all samples using the GATK's HaplotypeCaller in the "GENOTYPE_GIVEN_ALLELES" mode [235,236]. These variant calls were converted to a text based file format (**Table 3.4**) and allelic noise was annotated as an additional quality metric. Allelic noise was measured as the fraction of reads supporting the alternate allele relative to the total number of reads in all samples genotyped as homozygous for the reference allele by the HaplotypeCaller. Samples with called somatic variants were excluded from the calculations of allelic noise. If multiple alternate alleles were present, only the highest alleleic noise was recorded. Using allelic noise and the quality metrics annotated by the GATK's HaplotypeCaller, the overall quality of the variants was assessed manually. Validation of the highest quality variants was attempted and the results of the validation are shown in **Table 3.3** (run 2) and **Table 3.5**.

Table 3.4. Tissue-specific mosaic SNVs identified by MuTect.

Quality metrics for each mutation are shown (n = 284). The "Called" column indicates samples in which the mutation was identified by a germline variant caller. For each mutation, the total number of sequence reads (DP) and the number of reads supporting the alternate allele (AD) are reported for each individual. Attempted validation of variants by targeted sequencing (orange), or both pyrosequencing and targeted sequencing (green) are indicated. Samples from brain (B), heart (H), or kidney (K) are noted as in Table 3.1.

Identified in	Chrom	Position	Ref	Alt	In dbSNP	Map Q	MQ Rank Sum	BQ Rank Sum	RP Rank Sum	Fisher Strand Bias	Symmetric Odds Ratio
1349B	1	1431165	C	T	TRUE	56.52	-5.16	-1.82	3.01	0	0.04
4907B 5144H	1	6505917	A	G	TRUE	56.28	-5.76	3.3	-3.72	0	3.2
4999B	1	16357117	A	G	FALSE	56.31	-5.17	-2	-0.85	0	4.89
4671H	1	17085872	A	G	TRUE	51.13	-17.29	-9.54	-1.8	18.44	4.88
4671B	1	22304463	A	G	TRUE	35.61	-2.71	21.03	1.35	0	0.67
4999H	1	26671339	A	C	FALSE	58.97	-1.42	-8.33	3.4	9.15	6.99
5408B	1	26671658	G	A	TRUE	58.89	0.41	-5.78	-1.94	6.68	2.09
5115B	1	43166630	G	A	TRUE	58.97	-5.99	3.02	4.4	0	5.13
5144B	1	109792751	T	C	TRUE	61.56	-1.45	-2.68	1.48	0	0.76
5144H	1	120539742	G	A	TRUE	49	-6.92	2.54	2.82	4.79	5.35
1349K 4722H	1	120612043	G	T	TRUE	49.72	4.95	-2.02	3.65	9.26	1.62
5408H	1	145209116	T	C	FALSE	43.17	-5.25	1.63	-1.63	15.54	0.45
4899B	1	145209191	G	C	FALSE	49.79	-5	3.07	-4.58	0	0.68
4849B	1	151238513	C	T	FALSE	59.12	-0.28	-3.14	0.04	0	0.3

Identified in	Chrom	Position	Ref	Alt	In dbSNP	Map Q	MQ Rank Sum	BQ Rank Sum	RP Rank Sum	Fisher Strand Bias	Symmetric Odds Ratio
5408B	1	152278049	A	C	TRUE	33.71	5.93	10.41	1.94	19.73	4.63
797H	1	152883347	C	G	FALSE	58.81	0.37	-6.69	-2.9	25.25	4.29
5408H	1	155187144	A	G	TRUE	37.56	0.51	-3.06	1.02	0	4.17
5419H	1	155292776	G	C	FALSE	57.35	1.92	-12.48	6.04	73.23	8.63
5176H	1	155733245	C	T	FALSE	58.26	-12.47	-1.59	-12.83	6.32	0.25
4899H	1	162773301	G	A	TRUE	50.07	-5.58	2.15	-3.58	4.24	5.55
4907H	1	162775267	C	T	TRUE	49	1.98	-17.83	0.47	0.66	0.8
5176B	1	186276588	T	C	FALSE	53.36	3.03	-16.16	-2.48	0.76	0.87
5176B	1	196759282	C	T	TRUE	33.39	7.09	-1.59	0.84	8.47	2.92
4722H 797H	1	243333027	A	G	TRUE	44.15	-5.61	13.42	0.95	18.4	9.72
5408H	1	248085149	C	G	TRUE	49.98	-3.42	-1.3	1.63	2.93	0.29
4671H	2	27324340	G	A	TRUE	59.65				0	0.52
5144H	2	48757264	C	G	TRUE	56.29	-3.78	1.62	0.38	0	1.59
5176H	2	73901980	G	A	TRUE	59.07	-4.96	-4.53	-4.38	0	0.05
4899H	2	96619734	C	T	TRUE	54.7	-11.02	-4.22	1.94	0	0.91
4899B	2	96688929	G	A	TRUE	38.81	-10.11	5.39	-4.67	0	0.31
4899H	2	130897234	C	T	TRUE	50.7	1.23	6.5	2.4	1.23	0.8
4849B	2	130900091	C	T	FALSE	45.56	-1.49	-2.1	-0.37	0	1.06
5278K	2	132110862	G	C	FALSE	35.17	1.87	-1.02	1.13	67.46	4.09
4849B	2	132229694	C	T	FALSE	49.88	-2.94	-1.07	-1.63	0	0.68
5176H	2	132236921	G	A	TRUE	55.49	-13.16	9.83	4.05	2.55	5.12
4999H	2	171627348	G	C	FALSE	56.04	0.98	-2.45	2.68	86.15	5.54
4722B	2	171627363	A	C	FALSE	57.54	0.81	-13.9	-3.96	2.32	3.07
5419B	2	197187328	G	A	FALSE	59.53	-0.7	-2.26	0.91	0	0.36

Identified in	Chrom	Position	Ref	Alt	In dbSNP	Map Q	MQ Rank Sum	BQ Rank Sum	RP Rank Sum	Fisher Strand Bias	Symmetric Odds Ratio
5176H	2	198363501	C	T	TRUE	59.05	-1.69	-1.5	-1.56	0	1.43
5176B	2	200684266	C	A	FALSE	59.35	0.75	-6.4	1.53	0	3.3
5144B	2	207911121	A	G	FALSE	39.1	1.33	-1.49	-1.21	0	0.03
5176H	2	234590616	C	A	TRUE	58.89	-1.19	-1.74	0.67	6.35	0.83
1349B 4671B 5176B 5115B 4899H	3	10114944	A	C	TRUE	57.85	-17.26	4.84	4.46	2.17	4.96
4899H	3	42700630	A	G	TRUE	59.22	0.33	-14.02	4.48	2.04	0.51
4849B	3	49724808	T	C	TRUE	56.57	-1.81	-4.56	0.95	0	0.66
5176B	3	65425591	T	C	TRUE	60.33	-0.97	-2.31	1.2	0	0.88
5176B	3	65425594	C	T	FALSE	60.3	-0.49	-1.23	-0.41	0	1.4
5176B	3	123376027	T	C	FALSE	59.51	-1	-0.89	1.17	0	2.29
1349K	3	132441084	C	T	FALSE	59.23	0.8	0.9	2.11	0	0.42
4671B	3	195400728	A	G	TRUE	46.44	-9.22	-10.65	4.82	0	0.67
5419B 5115H	4	4239589	C	T	TRUE	51.24	-10.22	-8.02	-2.12	23.18	6.87
5419H	4	93225845	G	T	FALSE	59.68	1.13	-0.39	-1.94	0	1.54
4849H	4	141771717	T	G	FALSE	57.73				0	0.03
797H	4	147824864	G	A	TRUE	59.6	0.54	1.93	5.73	0	0.45
5408H	4	152024138	A	C	TRUE	47.81	-3.99	-4.65	-1.75	8.52	2.41
5403H	4	153870322	A	C	TRUE	55.37	5.64	11.99	3.86	0	0.64
5419B	5	94956735	C	T	FALSE	59.4	1.92	0.54	0.84	0	0.68
5115B	5	104043367	A	G	FALSE	59.84	-1.72	-1.71	-0.29	0	0.36
5115B	5	104043370	G	A	FALSE	59.82	0.47	-0.07	0.46	0	0.24
5278K	5	139781730	G	A	FALSE	59.58	0.42	1.86	0.42	0	0.57
4899H	5	139907772	T	G	FALSE	58.99	1.12	-25.47	4.23	27.56	9.37

Identified in	Chrom	Position	Ref	Alt	In dbSNP	Map Q	MQ Rank Sum	BQ Rank Sum	RP Rank Sum	Fisher Strand Bias	Symmetric Odds Ratio
5419H	5	140050940	C	T	TRUE	52.9	-7.36	-3.86	2.17	0	0.94
4907B	5	167988433	T	A	TRUE	58.16	-3.76	-2.02	0.36	2.53	0.13
5176B	5	178949876	C	T	FALSE	38.17	0.01	0.06	2.48	13.64	8.01
5176B	5	180431746	C	T	TRUE	56.24	-2.47	-0.45	1.23	0	1.43
5403H	6	9616237	G	A	TRUE	54.18	4.79	0.24	-2.51	0	1.02
5176H	6	10756728	C	T	TRUE	58.83	-1.37	2.05	0.04	0	2.22
5391B	6	16327903	C	A	TRUE	60	-1.23	-2.44	3.68	0	1.18
5144B	6	16327921	C	A	TRUE	60.18				0	0.3
4907H 5176H	6	27114569	T	C	FALSE	54.11	-5.12	2.85	-0.84	0	2.81
4999B	6	29857360	G	A	FALSE	55.8	-10.25	2.14	4.37	2.55	0.34
5403H	6	29912153	A	G	TRUE	53.99	-15.56	1.46	2.91	34.41	7.38
5176B	6	31963559	A	G	TRUE	31.18	5.77	3.23	1.68	3.03	7.16
4671H	6	31977552	C	G	TRUE	44.18	-5.53	0.97	-5.29	2.58	2.22
4899B	6	32713674	C	A	TRUE	55.25	-15.6	-5.28	9.36	11.14	2.02
4899B	6	32714091	A	G	TRUE	58.65	-10.36	-8.09	2.52	2.55	3.51
4899H	6	44221048	T	G	FALSE	58.23	-18.31	4.96	-2.09	5.09	6.77
5419B	6	51914982	C	T	FALSE	59.34	0.58	-2.02	0.08	0	0.44
4999H	6	130504496	G	A	TRUE	55.96	1.28	-0.24	-0.95	1.14	0.84
5419B	6	157100005	C	A	TRUE	59.4	0.21	5.46	0.45	0	2.64
5115B	6	170871061	G	A	FALSE	60.91	-1.09	-1.73	-0.92	0	0.3
5408B	7	1574061	A	C	FALSE	59.22	0.4	-22.23	4.02	25.02	6.45
5419H	7	56149352	G	A	FALSE	59.42	-0.62	1.61	0.32	0	0.22
5419H	7	72413896	A	G	TRUE	46.11	-6.94	6.02	-0.02	15.56	6.56
5144H	7	72414028	A	G	FALSE	40.97	-0.05	-2.66	1.18	0	0.76

Identified in	Chrom	Position	Ref	Alt	In dbSNP	Map Q	MQ Rank Sum	BQ Rank Sum	RP Rank Sum	Fisher Strand Bias	Symmetric Odds Ratio
4899H	7	72419616	C	A	FALSE	45.86	-11.73	-5.92	6.48	33.73	10.28
5278B	7	72420467	G	A	FALSE	49.94	-6.75	3.52	-4.97	0	1.09
5391B	7	75028240	T	C	TRUE	27.74	-4.46	0.78	2.49	3.06	0.25
5391B	7	75028241	G	A	TRUE	27.77	-4.79	-2.41	2.1	2.98	0.24
5408B	7	82583018	A	T	FALSE	58.98	-0.23	-1.96	0.69	0	0.46
4722B 4671H 5278K 5391B 5176H	7	99795228	G	T	TRUE	58.09	-15.22	-1.9	3.62	0	0.56
5408B	7	100028822	G	A	FALSE	59.48	-0.34	-1.19	1.85	0	0.44
4899H	7	141767225	C	T	TRUE	44.64	-2.16	5.14	0.7	0	0.55
4671B	7	145691990	T	C	FALSE	57.42				0	0.03
4899H	7	151970945	G	T	FALSE	47.37	-1.89	-0.78	-2.2	0	0.62
5115H	7	152513619	G	A	FALSE	36.73	4.44	10.01	0.8	21.59	10.49
5278B	7	153750140	G	A	FALSE	56.86	-6.1	5.65	-2.79	4.57	1.24
5176B	8	7274308	A	G	TRUE	30.32	11.79	7.99	2.38	102.35	8.47
5176H	8	11188922	T	C	TRUE	57.62	-16.25	-5.89	-0.69	0	0.55
4999H	8	143959174	T	C	TRUE	57.03	-8.62	0.79	0.3	6.55	0.14
5115B	9	20414343	A	G	FALSE	58.56	1.24	-0.49	-0.76	0	0.77
797H 4899H	9	34725438	A	G	TRUE	32.5	8.02	-8.52	-1.62	14.47	3.24
4671B	9	34834814	C	G	FALSE	46.99	-12.46	2.66	0.92	0	1.09
5176B	9	34834937	A	G	TRUE	42.72	-10.12	12.77	2.32	0.89	0.57
4899H	9	34835480	C	T	FALSE	47.69	-8.44	11.35	-0.16	5.19	8.38
5176B	9	66457019	G	A	TRUE	54.36	3.77	8.28	1.78	0	0.28
4671B 4999H	9	66457216	A	G	TRUE	49.85	-7.47	5.12	-0.79	0	0.79
5115B	9	78790138	A	G	TRUE	57.84	-7.5	-6.77	-6.02	0	0.93

Identified in	Chrom	Position	Ref	Alt	In dbSNP	Map Q	MQ Rank Sum	BQ Rank Sum	RP Rank Sum	Fisher Strand Bias	Symmetric Odds Ratio
5115B	9	78790143	G	A	TRUE	57.76	2.7	-5.46	-2.84	0	0.44
5403H	9	96439007	G	T	TRUE	55.1				0	0.59
5408B	9	130457371	A	G	FALSE	60.95	-3.01	0.39	-0.02	0	0.01
4849H	9	130457374	G	A	FALSE	60.97	-0.59	-2.17	-0.75	0	0
5408H	9	131020812	C	A	TRUE	59.37	0.67	-0.66	0.3	0	0.23
4899H	9	137742628	G	C	FALSE	56.7	3.4	-15.33	-5.24	13.46	5.82
4907H 4999B	10	21805480	T	C	TRUE	53.12	4.53	0.19	-0.19	2.22	1
5115B	10	21806056	G	A	FALSE	57.76	5.06	13.86	3.49	0	0.35
4671B 4907H	10	38647315	T	C	FALSE	55.99	-11.39	5.25	-1.12	0	5.43
5176H	10	52502717	G	A	TRUE	54.84	-6.71	4.27	0.65	0	3.34
797B	10	79397364	C	T	FALSE	58.77	-1.04	3.17	3.64	0	1.12
4849B	10	99330268	A	C	FALSE	59.49	1.56	-21.33	6.25	35.8	8.96
5419B	11	47600657	C	T	FALSE	59.32	-0.28	-1.09	1.5	0	0.55
4907H	11	49208319	G	A	TRUE	57.54	-15.57	-3.63	-0.18	1.69	1.2
4999B	11	55656479	T	G	FALSE	56.95	-5.9	-3.23	4.47	0	0.17
4671B	11	56468111	C	T	TRUE	48.76	-5.97	-1.38	3.05	5.56	6.02
4849H 4999H 5115B	11	89531764	T	C	TRUE	44.33	-5.47	-1.24	-2.25	0	0.84
4899H	11	95825221	C	T	FALSE	59.22	-0.61	-5.19	1.13	0	1.07
5176B	11	95825407	C	T	TRUE	60.14	-0.24	-1.29	-0.15	0	1.44
4907H	11	117073821	T	G	TRUE	59.49	-0.29	-19.22	4.44	18.99	8.5
5176B	11	117789345	G	C	TRUE	60.05	0.86	-1.43	-1.72	0	1.43
797B	12	2039173	A	G	TRUE	55.59	1.81	-4.88	0.16	0	0.03
4899H	12	9447029	T	C	TRUE	41.18	-3.99	-7.25	-1.48	24.63	8.85

Identified in	Chrom	Position	Ref	Alt	In dbSNP	Map Q	MQ Rank Sum	BQ Rank Sum	RP Rank Sum	Fisher Strand Bias	Symmetric Odds Ratio
4899H	12	9447061	C	T	TRUE	38.15	-0.22	6.39	0.92	4.88	5.9
5391H	12	9573223	C	A	TRUE	43.95	-2.95	-11.49	2.39	2.73	0.51
5176H	12	11506749	T	C	FALSE	48.26	-23.2	12.39	1.49	11.23	1.05
5408H	12	13153153	C	T	FALSE	57.89				0	0.06
797B	12	13153397	G	C	TRUE	51.96				0	0.05
5144H;797H	12	21623127	A	G	FALSE	43.19	-3.64	8.1	1.31	0	3.92
5278B	12	31244665	C	T	TRUE	47.89	-2.38	-1.99	-0.85	0	0.17
5278B	12	31244703	C	T	FALSE	47.42	-20.41	6.56	-7.61	24.36	1.45
5176H	12	31244784	C	A	FALSE	52.89	-6.05	1.89	-3.08	0	1.38
4849H	12	31255198	G	A	FALSE	51.71	-2.31	2.21	2.43	0	0.39
5419B	12	31255209	G	A	FALSE	52.3	-11.27	-5.84	5.06	0	4
5419B	12	49213447	G	A	FALSE	59.4	-0.01	-1.02	-1.13	0	0.13
5144B 5391B 5419H	12	49427679	C	T	FALSE	59.55	-1.3	-5.86	1.14	0	0.84
5115B;5176H	12	52699548	T	C	TRUE	55.52	-19.2	-29.46	-0.66	5.05	1.12
4671H;5278B	12	53298675	A	C	FALSE	56.13	-11.37	6.78	-3.27	9.54	2.02
4671H	12	53298699	G	A	TRUE	54.06	-11.75	5.55	7.39	6.64	2.27
4999H	12	53345343	G	A	FALSE	59.36	0	-2.03	-0.38	0	0.51
5403H	12	57112377	A	G	FALSE	58.74	0.22	-8.53	-0.28	4.25	1.42
5408B	12	63964599	T	C	FALSE	57.24	-15.63	0.68	-3.58	0	0.51
5408B	12	63964600	G	A	FALSE	57.21	-15.25	-7.48	-3.73	0	0.47
4671B	12	123950169	G	A	FALSE	59.46	0.32	1.49	0.33	0	1.24
5278K	12	132547087	G	A	TRUE	60.82	-1.07	-2.52	-0.01	0	0.77
4849B 5144B 5115B	12	132547096	G	A	TRUE	60.96	-1.74	-0.64	2.1	0	0.88

Identified in	Chrom	Position	Ref	Alt	In dbSNP	Map Q	MQ Rank Sum	BQ Rank Sum	RP Rank Sum	Fisher Strand Bias	Symmetric Odds Ratio
4671B	12	132589038	A	G	FALSE	32.25	3.72	3.57	1.59	9.24	6.09
5144B	13	20066994	T	C	TRUE	46.76	-14.92	9.26	-9.06	11.58	0.88
4899H	13	24895393	A	G	TRUE	42.14	-13.94	17.06	0.49	33.59	3.6
4722B	13	25016762	G	A	TRUE	58.27	-7.56	-6.26	-1.68	14.54	2.99
5419B	13	25021201	G	A	TRUE	49.16	-10.28	6.21	5.67	18.44	4.49
5176H	13	25052393	G	T	TRUE	57.78	-11.43	-5.04	3.92	4.57	0.89
1349K	13	32818263	G	A	FALSE	59.48	-0.2	2.53	0.12	0	1.07
5419B 4899B	13	45523879	T	C	FALSE	58.97	-3.25	-5.02	-2.72	0	0.01
1349K	13	113742714	G	A	FALSE	59.67	-0.1	-2.12	-0.54	0	0.51
4671B	14	19685516	C	T	FALSE	26.23	2.77	-1.87	-0.29	5.95	1.25
5278B	14	23744826	T	A	TRUE	61.91				0	0.04
5278B	14	23744829	C	T	TRUE	61.88	-1.04	1.64	0.43	0	0.14
4671H	14	70924335	C	T	TRUE	52.22	-5.2	3.5	-0.67	0	3.07
1349B 5391B 5176H	14	70924450	C	T	TRUE	49.77	-13.69	-10.78	-1.19	1	0.48
5176B	14	74008216	C	G	TRUE	55.13	-6.21	-3.24	1.07	0	3.39
5176B	14	74008309	C	T	TRUE	55.78	-7.8	-5.43	-0.08	4.12	1.34
5176B	14	106053577	A	G	TRUE	37.55	1.22	1.11	0.55	12.93	2.23
4849B	14	106436096	C	T	TRUE	59.24	-1.49	-1.58	0.28	0	0.34
5176B	14	106518415	C	A	TRUE	41.63	-1.94	-4.32	-0.56	0	1.4
5176H	14	106691691	A	G	FALSE	38.41	-2.6	-3.1	-0.09	0	0.32
4849B	15	20454011	G	A	TRUE	52.15	-5.71	-2.67	2.25	3.19	5.13
4722B	15	20462613	G	C	TRUE	55.01	7.23	-7.01	1.25	1.25	0.45
5408B	15	22413743	C	T	FALSE	42.91	-2.19	-1.22	0.67	0	0.6

Identified in	Chrom	Position	Ref	Alt	In dbSNP	Map Q	MQ Rank Sum	BQ Rank Sum	RP Rank Sum	Fisher Strand Bias	Symmetric Odds Ratio
4849B	15	22413810	G	C	FALSE	52.43	-5.72	0.4	-3.69	0	0.91
5403B	15	28370205	T	G	FALSE	59.33	0.44	-22.9	12.14	40.95	8.85
4722H	15	28474439	C	T	FALSE	44.63	1.6	2.63	0.09	0	0.02
5403H	15	28501283	T	G	FALSE	58.75	-0.71	-16.06	-1.59	15.32	6.39
4722B 5391B	15	29009175	G	A	TRUE	50.07	-7.08	2.85	5.02	10.76	7.22
4899H	15	29009243	G	A	FALSE	49.93	-9.7	-3.16	2.04	26.08	8.53
5403H	15	44487184	C	T	FALSE	59.59	-0.21	-2.14	1.45	0	0.47
1349K	15	50784955	C	A	TRUE	59.07	-14.28	-5.26	-1.12	0	0.77
5408H	15	90320161	G	A	TRUE	58.4	1.61	-0.54	-0.4	0	0.22
4899H	15	90320173	A	G	TRUE	58.53				0	0.06
5403H	15	100252741	A	C	TRUE	57.45	0.54	-0.58	1.14	0	3.78
5419H	16	426147	G	A	FALSE	59.27	0.06	1.91	2.29	0	0.69
5144B	16	1306918	C	T	TRUE	56.09	-11.56	10.25	3.24	0	0.7
5144B	16	1306921	G	A	TRUE	56.14	-11.5	2.04	1.94	0	0.71
5176H	16	4944518	C	G	TRUE	59.37	-2.15	4.23	0.32	0.74	0.79
5144H	16	28603714	C	T	FALSE	58.18	-4.07	-2.46	-3.47	0	1.2
5176H	16	32890639	T	C	TRUE	45.38	-6.44	-7.23	-9.81	21.67	3.58
797B	16	33738633	G	A	FALSE	47.98	-5.68	3.69	-1.26	3.3	5.42
4722H 5408H	16	65839631	T	C	TRUE	59.39	-1.58	-0.63	-1.7	0	1.42
5176H	16	67289691	G	A	FALSE	59.33	0.21	1.29	1.21	0	1.07
5115B	16	69996928	G	A	TRUE	37.7	-0.99	-1.35	2.86	0	3.4
5419B	16	70154480	A	G	TRUE	56.61	-20.75	3.47	-4.72	4.02	1.11
4671B	16	71103269	T	C	TRUE	40.42	-1.17	-1.05	0.68	0	4.37
5115H	16	71956529	C	T	TRUE	53.75	1.62	-0.6	0.39	0	2.12

Identified in	Chrom	Position	Ref	Alt	In dbSNP	Map Q	MQ Rank Sum	BQ Rank Sum	RP Rank Sum	Fisher Strand Bias	Symmetric Odds Ratio
5278K	16	72821609	G	A	FALSE	57.81	-1.41	-1.42	-1.43	0	1.47
5278B	16	72821624	G	A	FALSE	57.42	0.35	0.86	0.25	0	0.4
4999H	16	72992617	C	T	FALSE	59.15				0	0.61
4722B	16	74372644	A	G	FALSE	51.88	-10.38	-6.13	-1.27	3.01	1.91
4849B	17	7466636	G	A	FALSE	59.36	1.13	-1.6	-0.58	0	1.27
4899H	17	7843053	C	T	TRUE	59.5	0.18	-11.66	4.53	0	0.66
4671B 4899H	17	18682505	T	C	TRUE	41.83	-9.24	6.17	1.62	1.13	0.99
5115B	17	21199403	C	T	TRUE	55.7	-5.23	-1.51	0.48	0	0.62
4899H	17	21202191	C	A	TRUE	58.33	-11.99	2.01	1.88	2.87	7.13
4899B	17	39274087	G	C	TRUE	46.55	-9.29	-4.72	-3.1	9.15	0.81
5391B	17	39274416	C	T	TRUE	57.65	-15.4	-2.91	-10.47	1.69	0.26
5408H	17	39334314	A	T	TRUE	57.87	-4.38	-2.04	-0.06	0	0.9
5419B	17	43317923	G	A	FALSE	59.48	-0.22	-1.46	1.04	0	0.82
4899H	17	45127107	C	G	FALSE	45.18	-22.73	5.04	-1.18	9.04	7.76
5278B	17	45664677	C	T	FALSE	44.39	-6.57	12.56	5.02	10.7	9.12
5403H	17	56603934	A	C	TRUE	57.78	-1.99	1.7	-4.34	0	0.63
4899H	17	71443848	C	T	TRUE	59.48	0.33	3.03	0.48	0	1.5
4722H	17	74271953	C	T	FALSE	59.06	-0.59	0.72	0.77	0	0.54
5278B	17	78064060	G	A	FALSE	56.28				0	0.44
4899H	18	9887493	T	C	TRUE	59.3	2.38	-12.7	-3.34	0	0.68
4849H	18	14156352	A	C	FALSE	46.75	-3.75	5.54	2.58	2.6	4.98
4849B 797H	18	61391468	C	T	TRUE	51.95	2.43	-2.32	-0.26	0	0.93
4849B	19	1827059	A	G	FALSE	58.94	-1.28	-13.41	0.89	16.75	0.06
1349K	19	2643312	C	T	TRUE	53.45	-6.92	3.32	-4.34	4.08	3.48

Identified in	Chrom	Position	Ref	Alt	In dbSNP	Map Q	MQ Rank Sum	BQ Rank Sum	RP Rank Sum	Fisher Strand Bias	Symmetric Odds Ratio
5419H	19	10752079	A	C	FALSE	58.89	-0.68	-0.13	0.3	0	0.17
4999B 5419H	19	10752080	C	A	FALSE	58.94	2.23	-2.57	-5.61	1.91	0.24
5115B	19	12772090	C	T	TRUE	59.51	-1.37	10.53	0.64	2.36	1
5408B	19	14884843	C	T	TRUE	55.63	-2.81	-1.27	0.63	0	1.56
5144B	19	16640583	C	T	FALSE	59.48				0	0.21
4999B	19	17006736	A	G	TRUE	55.75	-1.74	-6.33	4.57	0.79	0.77
5115H	19	17397477	G	T	TRUE	57.31	1	-3.41	-0.23	3.77	1.59
4999B	19	18879426	A	C	FALSE	58.81	0.41	-14.85	3.24	61.8	7.06
5115B	19	22363844	G	A	TRUE	53.69	-0.39	-14.77	-16.75	11.75	1.43
5391B	19	22585603	T	C	TRUE	53.54	-8.98	3.9	-5.62	0	0.91
4907H	19	22836805	G	A	TRUE	52.5	-18.77	-9.14	-12.84	0	0.58
5391H	19	33490585	G	A	TRUE	49.39	-9.15	-6.29	-0.73	6.65	1.49
4849H 4907H 4899B	19	33517507	C	T	FALSE	58.84	-7.62	-1.09	-5.16	2.96	4.14
5391B 5403B	19	40376323	A	G	FALSE	34.51	-2.29	1.68	1.34	2.36	1.73
4849H	19	41627496	C	T	TRUE	55.88	-3.69	-1.3	0.83	0	0.84
5144B	19	43031434	T	A	TRUE	58.51	-15.03	1.56	-7.05	8	2.76
4999H	19	48305574	A	G	FALSE	58.88	0.12	-11.63	1.66	7.19	1.75
4849H	19	48305694	A	G	TRUE	58.6	2.16	-11.48	-1.14	4.95	0.3
797B	19	49926533	G	C	TRUE	59.52	-0.44	-0.53	2.51	0	0.64
4907H	19	50463670	T	G	TRUE	46.76	-1.34	22.54	6.24	3.99	1.07
4671H	19	51274851	A	C	TRUE	53.42	-16.19	7.99	5.29	25.02	2.38
4899B	19	52132668	T	C	TRUE	42.28	-0.72	3.64	2.87	12.07	3.15
5176H	19	53303332	C	T	TRUE	58.7	-5.1	1.92	-2.97	0	4.61

Identified in	Chrom	Position	Ref	Alt	In dbSNP	Map Q	MQ Rank Sum	BQ Rank Sum	RP Rank Sum	Fisher Strand Bias	Symmetric Odds Ratio
4849B 5408H 5176B	19	53770304	T	G	TRUE	42.42	-4.35	-0.51	-1.79	2.3	4.45
4899B	19	53788121	A	G	TRUE	49.4	-3.1	-2.04	0.64	0	3.23
4907B	19	54723995	G	C	TRUE	43.21	-13.11	-10.04	-1.05	92.34	8.12
4849H	19	54726237	C	G	TRUE	30.03	4	0.72	0.9	8.75	6.18
5176B	19	54744128	A	T	FALSE	43.38	-2.91	-1.33	-2.22	0	0.11
5176B	19	54744133	G	A	TRUE	43.49	-4.9	5.71	-0.42	0	0.27
1349B 4671H	19	55324635	T	C	TRUE	53.41	-5.97	2.48	5.28	0	0.44
5176H	19	58371368	T	G	TRUE	54.4	-6.46	4.75	2.4	0	5.76
4671B	19	58421080	G	T	TRUE	29.96	12.98	-12.39	-0.98	4.75	1.09
4907H	20	1585397	T	C	TRUE	42.92	-8.36	7.92	4.44	1.08	0.8
4907H	20	1585446	G	A	TRUE	39.09	-7.53	7	-2.17	7.34	1.6
5419H	20	26061803	C	A	TRUE	52.63	-13.89	-6.68	2.22	11.09	4.89
5176H	21	11098729	G	A	FALSE	57.57	-0.27	-1	-2.21	0	0.74
5419H	21	36410912	G	C	TRUE	47.25	6.92	-2.52	5.29	0	0.38
4899H	21	46011298	G	A	TRUE	45.83	-4.94	-3.83	-0.86	15.81	7.17
5278B	21	46011397	A	G	TRUE	51.95	-6.51	3.81	3.69	6.78	2.56
5278B	21	46011400	G	A	FALSE	51.83	-5.71	1.27	2.98	6.82	6.09
5278K	22	39357586	T	C	TRUE	41.58	-2.88	5.63	0.93	0	4.46
5391H	22	39811081	G	A	TRUE	59.56	-0.47	1.73	0.87	0	0.9
4671H	22	42523636	C	A	TRUE	56.47	-21.28	-10.15	-1.1	22.58	9.88
4671H	22	42908976	G	A	TRUE	52.59	-5.21	1.5	1.58	3.6	1.73
1349K	22	42911257	A	G	TRUE	38.72	-1.13	-9.96	3.38	103.89	9.75
4849H	22	50315971	C	G	TRUE	57.85				0	0.07

Identified in	Chrom	Position	Ref	Alt	In dbSNP	Map Q	MQ Rank Sum	BQ Rank Sum	RP Rank Sum	Fisher Strand Bias	Symmetric Odds Ratio
4849H	22	50315973	G	C	TRUE	57.88	-0.68	8.32	5.32	1.31	0.84
4722H	X	18660190	C	T	FALSE	59.33	0.73	-2.63	-0.34	0	1.13
5115H	X	37028645	G	C	FALSE	56.12	3.43	-9.96	-1.02	6.13	2.35
5176H 5115H	X	51075901	G	A	TRUE	54.11	-13.66	-9.36	0.98	23.79	7.85
4907H	X	53652838	G	A	TRUE	55.76	-5.77	2.86	1.94	15.29	1.27
4907H	X	53652846	G	T	TRUE	56.22	-5.58	-3.09	-0.87	22.85	0.92
4671H 4907H	X	53652852	T	G	TRUE	56.88	4.17	4.21	-1.83	20.23	2.03
5408H	X	53966821	T	A	FALSE	59.46	-0.65	-0.79	2.7	0	1.03
5144B 797H	X	64956699	G	A	TRUE	57.73	-1.61	-0.8	1.15	0	0.53
5144B 797H	X	64956743	A	G	TRUE	58.03	-1.09	0.63	0.3	0	1.44
4671B	X	110987993	A	T	FALSE	59.16	-0.04	-1.59	-1.7	0	0.67
5144H	X	114425196	G	A	TRUE	45.36				0	0.39
5419B	X	136649009	C	T	FALSE	59.41	-0.14	2.01	0.54	0	1.11
Validation Legend:											
Targeted Sequencing											
Pyrosequencing and Targeted Sequencing											

Table 3.5. Pyrosequencing of potential tissue-specific variants in paired samples.

Well	Assay	Sample ID	Var. Pos.	Quality	Warnings	A Frequency (%)	C Frequency (%)	G Frequency (%)	T Frequency (%)
C1	NDUFS3	5419B	R	Passed		1.49	-	98.51	-
C2	NDUFS3	5419H	R	Passed		1.59	-	98.41	-
C3	NDUFS3	5419K	R	Passed		1.33	-	98.67	-
C4	NDUFS3	5176B	R	Passed		1.66	-	98.34	-
C5	NDUFS3	5176H	R	Passed		1.83	-	98.17	-
C6	NDUFS3	5176K	R	Passed		0.75	-	99.25	-
C7	Control	Control	Y	Passed		-	53.73	-	46.27
C8	Control	Control	Y	Passed		-	52.51	-	47.49
A1	GRID2	5419B	K	Passed		-	-	96	0
A2	GRID2	5419H	K	Passed		-	-	95	2
A3	GRID2	5419K	K	Passed		-	-	81	1
A4	PHKG1	5419B	K	Passed		-	-	96	4
A5	PHKG1	5419H	K	Check	Uncertain due to low peak height.	-	-	91	9
A6	PHKG1	5419K	K	Passed		-	-	94	6
A7	SLC9A5	5716B	Y	Passed		-	99	-	1
A8	SLC9A5	5716H	Y	Passed		-	99	-	1
B1	SLC9A5	5716K	Y	Passed		-	98	-	2
B2	QRICH2	4722H	R	Passed		3	-	97	-
B3	QRICH2	4722K	R	Passed		3	-	97	-
B4	TAB1	5391B	R	Failed	Failed due to low peak height. Uncertain surrounding reference sequence pattern.	9	-	91	-
B5	TAB1	5391K	R	Check	Uncertain surrounding reference sequence pattern. Uncertain due to low peak height.	7	-	93	-
B6	SNRNP35	4722H	R	Passed		2	-	98	-

Well	Assay	Sample ID	Var. Pos.	Quality	Warnings	A Frequency (%)	C Frequency (%)	G Frequency (%)	T Frequency (%)
B7	SNRNP35	4722K	R	Passed		2	-	98	-
C7	Control	Control	Y	Passed		-	47	-	53
C8	Control	Control	Y	Check	Uncertain due to low peak height.	-	47	-	53

For visualization, the properties of variants including "BaseQRankSum", "FS", "MQ", "MQRankSum", "ReadPosRankSum" and "SOR" were examined in the Illumina Platinum Genomes (see below) and variants called by MuTect. These properties were collectively scaled, principal components analysis was performed and the original variant features were transformed into the principal components.

3.2.3 In silico mixing experiment using NA12878 and NA12882

200x sequence data from NA12878 and NA12882 were downloaded from EBI (ERP001775). The specific runs chosen were ERR174324, ERR174325, ERR174326, ERR174327, ERR174328, ERR174329, ERR174330, ERR174331, ERR174332, ERR174333, ERR174334, ERR174335, ERR174336, ERR174337, and ERR174338 for NA12878 and ERR174347, ERR174348, ERR174349, ERR174350, ERR174351, ERR174368, ERR174369, ERR174370, ERR174371, ERR174372, ERR174373, ERR174374, ERR174375, ERR174376, and ERR174377 for NA12882. Sequence data were aligned to the 1000 Genomes phase 2 human reference genome using BWA MEM version 0.7.9a and aligned sequence reads were sorted using SAMtools [134,234]. Aligned sequence data were then combined into single files and *in silico* mixing with subsampling was performed using submixbam (https://github.com/DonFreed/submixbam) version 290fda over GIAB high-confidence regions (http://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/GIAB_integration/union13callableMQonlymerged_addcert_nouncert_excludesimplerep_excludesegdups_excludedecoy_excludeRepSeqSTRs_noCNVs_v2.19_2mindatasets_5minYesNoRatio_AddRTGPlatGenConf_filtNISTclustergt9_RemNISTfilt_RemPartComp_RemRep_RemPartComp_v0.2.bed.gz; accessed Oct 14th 2015; currently available from ftp://ftp-

trace.ncbi.nih.gov/giab/ftp/data/NA12878/analysis/GIAB_integration/). A variant calling and pipeline was written using Snakemake [237]. This pipeline called variants from mixtures using the GATK's HaplotypeCaller version 3.4-46. The sensitivity of the HaplotypeCaller for variants known to be present in NA12878 and absent from NA12882 was then evaluated using hap.py (https://github.com/Illumina/hap.py).

3.2.4 Simons Simplex Collection Analysis

Analysis of the data in the Simons Simplex Collection made use of cloud computing via Amazon Web Services (AWS) for variant calling (GATK HaplotypeCaller), merging gVCFs (GATK MergeGVCFs) and genotyping (GATK GenotypeGVCFs). Starcluster (http://star.mit.edu/cluster/) was used for deployment and configuration of clusters of virtual machines on AWS Elastic Cloud Compute (EC2) and a customized Amazon Machine Image (AMI) was created containing GATK 3.5-0, samtools 1.2, Python 3.5.1, and nda_aws_token_generator version 20b72 (https://github.com/NDAR/nda_aws_token_generator) [134,235,236]. c3.xlarge, r3.xlarge and r3.2xlarge instances were used for variant calling, merging gVCFs and genotyping, respectively. During variant calling and merging, the available disk space on each node was used as a complex resource to aid in job allocation. With c3.xlarge instances, ephemeral storage partitions were combined into single logical volumes using RAID 0. During genotyping, node ephemeral disk partitions were combined into a single network attached storage volume using GlusterFS (https://www.gluster.org/).

3.2.5 Simons Simplex Collection variant discovery

Aligned whole-exome sequence data from 8,950 individuals in the SSC was accessed through the National Database for Autism Research (NDAR) on Amazon Web Services

Simple Storage Service (AWS S3) (https://ndar.nih.gov/study.html?id=334). We excluded 16 individuals from families 11366, 11368, 11377 and 11380 due to data processing issues. Variants were called using the GATK (v. 3.5-0) HaplotypeCaller in gVCF mode with standard variant annotations and additional arguments -ploidy 5, -A GCContent and –A AlleleBalance over NimbleGen EZ-SeqCap v2.0 targets with 50 bp of padding [235,236]. gVCFs of 20 families were combined using GATK MergeGVCFs resulting in 120 merged gVCF files. All gVCF files were genotyped across capture regions in parallel using the GATK GenotypeGVCFs command with the arguments -stand_call_conf 25.0, -stand_emit_conf 20.0 along with the arguments used with the HaplotypeCaller as described above. The genotyping step had high memory requirements over some target regions, causing some jobs to fail even with 116 GB of memory allocated to the java virtual machine. Failed capture regions were repeated with the additional argument --max_alternate_alleles 5. However, we excluded 53 capture targets due to persistent memory errors (**Table 3.6**). These capture targets were highly enriched for overlap with known simple repeats (UCSC Simple Repeats Track in BED format; tested using BEDtools fisher; Fisher's Exact Test, $p < 0.00001$). Variant calls over each capture target were then concatenated and duplicate calls due to overlapping padded targets were removed.

Table 3.6. Capture targets excluded due to extremely high memory usage.

Chrom	Start	Stop
1	6192855	6192955
1	25362447	25362597
1	38332122	38332223
1	207237096	207237196
2	119748137	119748237
2	192225356	192225455
3	8671331	8671431
3	38355217	38355432
3	38627181	38627532
3	75790760	75790887
3	186302218	186302378
3	195505661	195518368
4	82013523	82013623
4	88534936	88537720
5	149583229	149583342
5	153085256	153085627
5	167995650	167996003
6	31380102	31380237
6	32485472	32485572
6	32546824	32546924
7	286376	286477
7	123190523	123190640
7	151552452	151552633
8	10464404	10470856
9	8331581	8331736
9	136637074	136637174
10	3208421	3208572
10	12142168	12142269
10	135103327	135103473
10	135438602	135439108
11	1015761	1018770
11	10823595	10823756
11	60778524	60778624
11	71238346	71238844
11	85979497	85979603
11	123454980	123455080

12	53207380	53208064
14	39783923	39784023
14	81574707	81574807
14	81574893	81575030
14	92537278	92537397
15	52548826	52548926
16	9017071	9017270
16	33965500	33965600
16	89178499	89178654
17	2297569	2297669
19	5787071	5787182
19	48613708	48613808
20	54963182	54963282
20	56236685	56237090
21	40717071	40717200
21	42551103	42551555
22	32827317	32827417

3.2.6 Simons Simplex Collection variant filtration

In addition to the variant annotations produced by the GATK, raw variants were annotated with the number of sequencing reads supporting the reference allele relative to total number of sequence reads. This information was added to the VCF's INFO field as the annotation "AbHetUser". Variants were filtered using the GATK variant quality score recalibration pipeline. The recommended parameters for whole-exome sequencing were used minus the –an QD parameter and with the additional parameter –an AbHetUser. These parameters were chosen for their superior sensitivity and specificity for validated *de novo* variants in the SSC. SNPs were filtered with a sensitivity tranche of 99.3% while indels were filtered with a sensitivity tranche of 98%.

3.2.7 Simons Simplex Collection de novo and mosaic variant identification

De novo variants were identified using the tool find_denovo, a tool we wrote in the C programing language, with default parameters (https://github.com/DonFreed/find_denovo). find_denovo identifies alleles which are present in children but absent from their parents. It then applies a number of filters including a minimum number of reads for all trio members (20), a minimum number of reads supporting the alternate allele in the child (3), a minimum phred-scaled confidence for the presence of the *de novo* allele in the child (20) and the absence of the *de novo* allele in the parents (20), and a maximum number of individuals genotyped for the allele in the cohort (2). *De novo* variant effects were then annotated using SnpEff [238]. Families 11060, 11431, 11628, 11714, 11905, 12173, 12230, 12401, 12456, 12809, 12879, 13143, 13949, 14025, and 14355 were excluded as more than 10 *de novo* mutations were observed in at least one child in the family.

Mosaic variants were identified from *de novo* variants using the binomial test to examine the alternative hypothesis that the *de novo* allele is supported by significantly fewer sequence reads than expected from the read depth. We use $p = 0.5$ as the expected fraction of sequence reads supporting the *de novo* allele. p-values were adjusted using the Benjamini-Hochberg procedure with a FDR of 0.05 and variants with $q < 0.05$ were called mosaic. In the final callset we add the requirement that mosaic variants must have an AARF of less than 34%. In addition, mosaic variants that were identified uniquely in our callset and not in the callsets produced by Iossifov *et al.* or Krumm *et al.* were filtered.

3.2.8 Phasing of variants in the Simons Simplex Collection

Variants identified as *de novo* in the Simons Simplex Collection were phased to nearby inherited variants to validate mosaic status and to determine the parental haplotype of the variant allele using phase-mosaic, a tool we wrote in Java and Python (https://bitbucket.org/donald_freed/phase-mosaic, version f47bcd). For each identified *de novo* variant, sequence data 500bp upstream and downstream of the variant was downloaded to the local machine from AWS Simple Storage Service (S3) for each member of the pedigree. Variants were then recalled using the GATK version 3.5-0 compiled with the VariantReadIds annotation. Phasing was then performed on the resulting VCF files.

3.2.9 Rates of mutation in the Simons Simplex Collection

Regions of 40x coverage were defined for each individual in quad families using BEDtools genomecov –bga with the resulting BedGraph file converted to a BED file using a custom script [239]. BEDtools was then used to intersect the 40x BED file for each member of a trio and the target capture file to produce a joint 40x BED file for the trio. Variants in the callset were annotated based on their presence or absence in the joint 40x region using

custom scripts. The length of the genome present in the joint 40x region was recorded for each child in a quad family. Finally, the rate of *de novo* mutation for each individual and each class of mutation was calculated from the size of the joint 40x region and the number of mutations in joint regions identified in the child. These rates were then extrapolated to the entire capture region.

Iossifov et al. previously reported a model of *de novo* variation in which siblings have a baseline rate of *de novo* mutation while probands have the same baseline rate and additional mutation due to their affected status [212]. We expand this model to distinguish between germline *de novo* and mosaic variation while incorporating errors in classification of mosaic status. In siblings the observed rate of mosaic or germline *de novo* variation was modeled as the sum of correctly and incorrectly classified baseline variation. In probands the models included correctly and incorrectly classified contributory variation in addition to the baseline variation. Classification error rates were for siblings modeled as either incorrectly classified baseline variation over correctly and incorrectly classified baseline variation. For probands, classification error rates were modeled as incorrectly classified baseline and contributory variation over correctly and incorrectly classified baseline and contributory variation.

These models were solved to obtain the rate and fraction of contributory variation using the observed rates of mutation and classification errors as measured by phasing validation. Classification error rates were calculated separately for probands and siblings and for germline *de novo* and mosaic classification. Uncertainty in classification error rates was modeled using the beta-binomial distribution with phasing validation results as model

parameters. A 95% credible interval was obtained through 10,000 permutations with classification error rates obtained by random draws from their respective distributions.

3.2.10 Simons Simplex Collection variant conservation

Variants were annotated with PhyloP conservation score, taxonomic conservation as reported by NCBI's HomoloGene database, and the probability of null mutations being deleterious ("pNull") as reported by ExAC [240-242]. BigWig files containing genome-wide PhyloP scores were downloaded from UCSC (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP100way/hg19.100way.phyloP100way.bw; accessed Nov. 10th, 2015) and were used to annotate variant conservation. Gene-level conservation was annotated by querying the NCBI's HomoloGene database using Biopython and Entrez to find the earliest taxonomic unit reported to sharing the gene containing the mutation [243]. These taxonomic units were then converted to numeric scores where 0 corresponds to conserved in Homo while 31 corresponds to conserved to the root of the HomoloGene taxonomic tree. ExAC gene summary data were downloaded from (ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/functional_gene_constraint/README_fordist_cleaned_nonpsych_z_data_pLI_2016_01_13.txt; accessed Feb. 11th 2016) and variants were annotated with the reported probability of their respective gene being intolerant of loss-of-function mutation. Using these data, mutations present in probands were compared to mutations present in siblings with the Wilcoxon rank sum test.

3.2.11 Gene target overlap and recurrence

Methods for analysis of gene target overlaps and recurrence were adopted, with modification, from Iossifov *et al.* [212]. RefSeq genes were downloaded from the UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables; accessed Jul. 16th 2015). The

"chr" prefix was removed from the chromosome names and the raw table was sorted by chromosome and position. Coordinates of coding sequence starts and stops were extracted from the RefSeq table in BED format using custom scripts and overlapping coding sequences were merged using BEDtools [239]. This file was intersected with the BED file of the target capture region and the length of each gene in the target region was calculated. These data were then combined with data of gene membership in gene sets from Supplementary Table 7 of Iossifov *et al.* and the high-quality callset to produce a table describing the number of observed mutations in each gene and each gene's set membership [212].

Given their observed contribution to ASD diagnosis, only mosaic missense and germline *de novo* LGD mutations were analyzed and these mutations were analyzed in both probands and siblings. These analyses were performed using a null length model where the probability of a mutation occurring within a gene is proportional to its length targeted for exome capture relative to the total size of the capture target. For every mutation-type, individual combination, we calculate the following: (1) The expected number of recurrent mutations and a p-value for the observed number of recurrent mutations from 10,000 simulations using sampling with replacement. (2) For each gene set from Iossifov *et al.* we calculate the expected number of genes harboring mutation present in the gene set, given the length of capture targets of genes within the set relative to the total length of all gene capture targets. Using a two-sided binomial test, we test for observed enrichment or depletion from the expectation based on the null length model.

For testing the enrichment of mosaic missense and LGD mutations in genes implicated in ASD, we used the approach described above with the target gene set of 107 candidate genes identified by De Rubeis *et al.* [210].

3.2.12 Pyrosequencing

Amplification and sequencing primers were designed for all loci using PyroMark™ software and the NCBI's Primer-BLAST [244]. Additionally, primers were checked for overlap with common SNPs using the UCSC Genome Browser [245]. Samples were amplified according to protocols in the Qiagen Pyromark PCR kit with a single biotinylated primer. Pyrosequencing was performed and data were analyzed by the Johns Hopkins Genetic Resources Core Facility.

3.2.13 Amplicon-targeted sequencing

Sequence libraries were generated from purified DNA according the Nextera XT DNA Library Preparation Guide. Sequence data were then generated on an Illumina MiSeq using a MiSeq Reagent Kit v2. Sequence reads were aligned to the human reference genome (UCSC hg19) using BWA and reads supporting the reference or alternate alleles were counted [234].

3.2.14 Sanger Sequencing

Mosaic variants and germline *de novo* variants for validation were chosen at random from variants present in samples on hand. In total 97 variants were chosen for validation, 50 germline *de novo* variants and 47 mosaic variants. Primers for polymerase chain reaction amplification were designed using Primer-BLAST [244]. Amplification was performed using DNA isolated from whole blood and Sanger sequencing was performed at the Johns Hopkins University School of Medicine Synthesis and Sequencing Facility.

3.3 Results

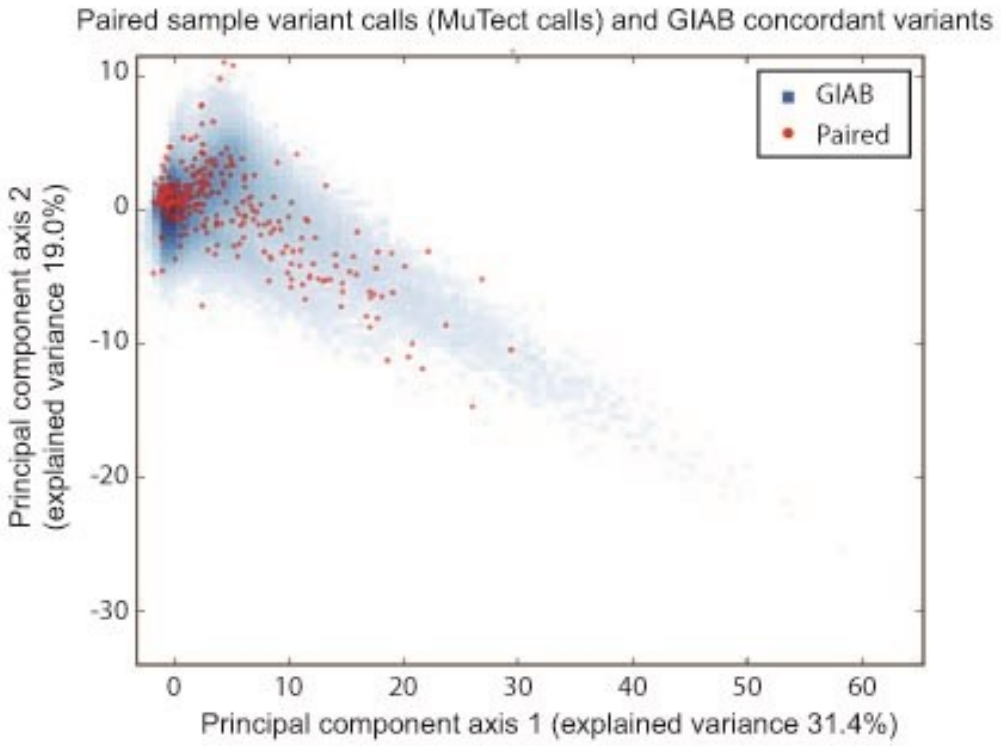
3.3.1 Tissue-specific mosaic mutation

Obligatory somatic mutations typically occur in a localized fashion in tissues that share a common developmental origin. To examine the contribution of tissue-specific mutations in ASD, we generated whole-exome sequence data from paired postmortem frontal cortex (n = 16) and heart (n = 14) or kidney (n = 2) samples from individuals diagnosed with ASD (n = 12) and controls (n = 4). Sequence data were generated using the Illumina HiSeq platform with an average sequence depth of 95x across capture targets (**Table 3.1; Table 3.2**). Variants were detected using the somatic variant callers Strelka and Mutect [132,133]. These programs detect mutations unique to single tissues from paired samples. Analyzing frontal cortex/heart or frontal cortex/kidney pairs resulted in the identification of 373 mosaic variants in the 32 samples. However, validation experiments indicated that all potential mutation loci were homozygous for the reference allele (*i.e.* all mutations chosen for validation were false positives; **Figure 3.1**; **Tables 3.3 to 3.7**). Our findings agree with previous results that tissue-specific mosaic mutation in brain rarely occurs at the level of detection afforded by standard whole-exome sequencing experiments [246].

Figure 3.1. Quality metrics of GIAB variants and MuTect calls.

(A) Principal components of scaled variant features in the GIAB callset visualized by log-transformed density (blue) overlaid with variants called by MuTect (red). Much of the density of GIAB variants was clustered around (0,0), with the MuTect variants having significantly more spread. 96.0% of GIAB variants cluster within a Euclidian distance of three from the origin, as did only 41.5% of variants called by MuTect. (B) Noise measurement in the GIAB and MuTect variants (log-scale). MuTect variants were called at sites with drastically more noise. While only 0.3% of GIAB variants had an allelic noise above 0.01, 60.2% of MuTect variants did.

A


Paired sample variant calls (MuTect calls) and GIAB concordant variants
GIAB
Paired
Principal component axis 2 (explained variance 19.0%)
Principal component axis 1 (explained variance 31.4%)
10
0
-10
-20
-30
0
10
20
30
40
50
60

B

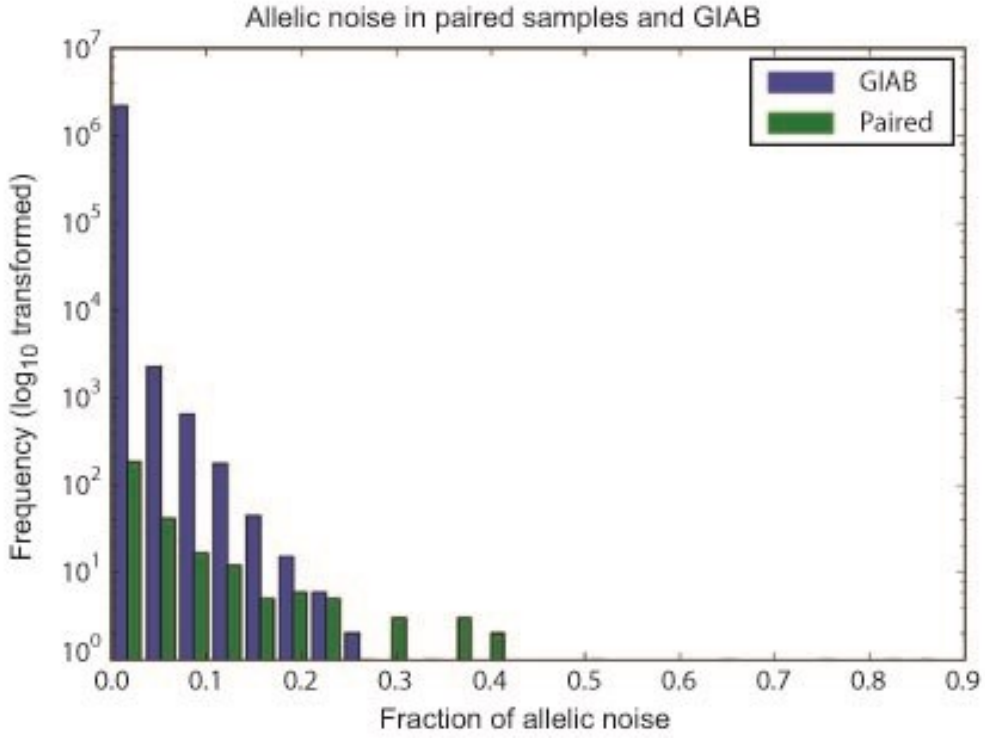

Allelic noise in paired samples and GIAB
GIAB
Paired
Frequency (log_{10} transformed)
Fraction of allelic noise
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Table 3.7. Tissue-specific mosaic indels identified by Strelka.

Attempted validation of variants by targeted sequencing (orange) is indicated.

Brain Bank ID	Chromosome	Position	Reference	Alternate	Secondary_Tissue: Ref;Alt	Frontal_Cortex: Ref;Alt
1349B	20	18123040	TTG	T	66;0	41;5
4671B	4	15987514	A	AGT	28;0	43;16
4671B	4	15987515	A	ATGCT	30;0	48;16
4671B	6	160132143	CCACTCACATTGCTCTGAT	C	95;0	46;4
4907B	7	74191612	GGT	G	84;0	51;6
5144B	7	116955170	ATCCTT	A	84;0	69;3
5144B	11	66411363	AGCTGCTGCTGCAGCAGCAGCC	A	156;0	169;4
5391B	1	241846828	CTG	C	66;0	42;6
5391B	19	17918776	A	AC	118;0	105;10
5408B	GL000220.1	119444	GCTCTCGCT	G	167;1	164;5
4722H	16	89299768	G	GCAAAT	93;0	49;6
4722H	17	79866732	G	GC	83;0	36;3
4722H	GL000220.1	131171	T	TTCTCTCTGTC	150;0	114;8
4671H	7	21659554	A	ATTAAT	70;0	67;3
4671H	8	10467680	TTCCTTC	T	331;0	461;14
4849H	1	172387390	ATTTTG	A	120;0	96;4
4849H	17	1585378	AACCACC	A	81;0	62;3
5144H	1	16376516	CAT	C	15;0	0;5
5278H	20	16362789	GGA	G	59;0	32;8
5391H	9	116184795	ATGTTT	A	73;0	54;3
5408H	9	116184795	ATGTTT	A	76;0	63;4

Brain Bank ID	Chromosome	Position	Reference	Alternate	Secondary_Tissue: Ref;Alt	Frontal_Cortex: Ref;Alt
4899B	6	160132562	CCACTCACATTGCTCTGAT	C	189;0	224;18
4899B	9	139910402	CAG	C	97;0	45;4
4899B	19	39220200	G	GGC	26;0	19;7
5176B	1	248616704	CTGCTGCG	C	335;1	97;4
5176B	9	138899193	GCA	G	72;1	38;21
5176B	9	138899204	GAC	G	72;0	34;19
5176B	9	138899222	TAC	T	57;0	31;10
5176B	16	30982808	ATCC	A	63;0	20;4
5176B	GL000220.1	120208	GCTGCTGCCTCTGCCTCCACGGTT	G	161;0	138;5
797H	14	20146543	A	AGTCCC	257;0	144;4
797H	16	4796952	TGAG	T	66;0	36;6
797H	16	24574672	TCTGGC	T	92;0	72;3
4899H	1	248616704	CTGCTGCG	C	419;0	152;13
4899H	9	135962590	CTGT	C	26;0	5;10
4899H	9	135962599	C	CG	30;0	6;9
4899H	9	135962600	C	CT	31;0	7;9
4899H	19	53418429	A	ACTTCT	73;0	82;6
5115H	3	53145933	GAAAAC	G	95;0	73;3
5115H	9	35364644	ATG	A	38;0	7;4
Validation Legend:						
Targeted Sequencing						

3.3.2 Detection of mosaic mutations from single samples

In next-generation sequencing data, reads supporting the alternate allele at variant sites are known to be under-represented due to biases against non-reference alleles [247]. Further, many variant callers explicitly assume a diploid model. Therefore, the extent to which existing germline variant callers accurately genotype mosaic mutations in unpaired samples is uncertain. To evaluate our ability to discover mosaic mutation occurring in single samples, we obtained the Illumina Platinum Genomes sequence including NA12878, an individual for whom a high-confidence callset is available (ERP001960) [248]. We then characterize the sensitivity of the GATK HaplotypeCaller for mosaic variants through *in silico* mixture experiments (**Figure 3.2**). For this experiment, we utilized 200x sequence data from the Illumina Platinum Genomes (ERP002490; See Materials and Methods). Sequence reads from NA12878 were mixed with sequence reads from her son NA12882 over regions known to harbor variants from the high-confidence GIAB callset. Mixtures were then subsampled to depths of 30x and 50x with random fractions of reads from N12878 and NA12882. Variants were then called from the mixture and the sensitivity of the variant caller was assessed for variants known to be present in NA12878 but not NA12882. Sensitivity was also assessed with different values of –ploidy argument which alters the expected AARFs of heterozygous variants. These results demonstrated that higher –ploidy settings improved sensitivity for low frequency mosaic variants at the cost of higher memory usage and longer runtimes (**Figure 3.3**).

Figure 3.2. Performance evaluation of submixbam.

(A and B) Evaluation of subsampling functionality of submixbam. Target depths are consistent with measured depth across a range of target depths. (C and D) Evaluation of mixing functionality of submixbam. submixbam performed highly precise and accurate mixing and subsampling.

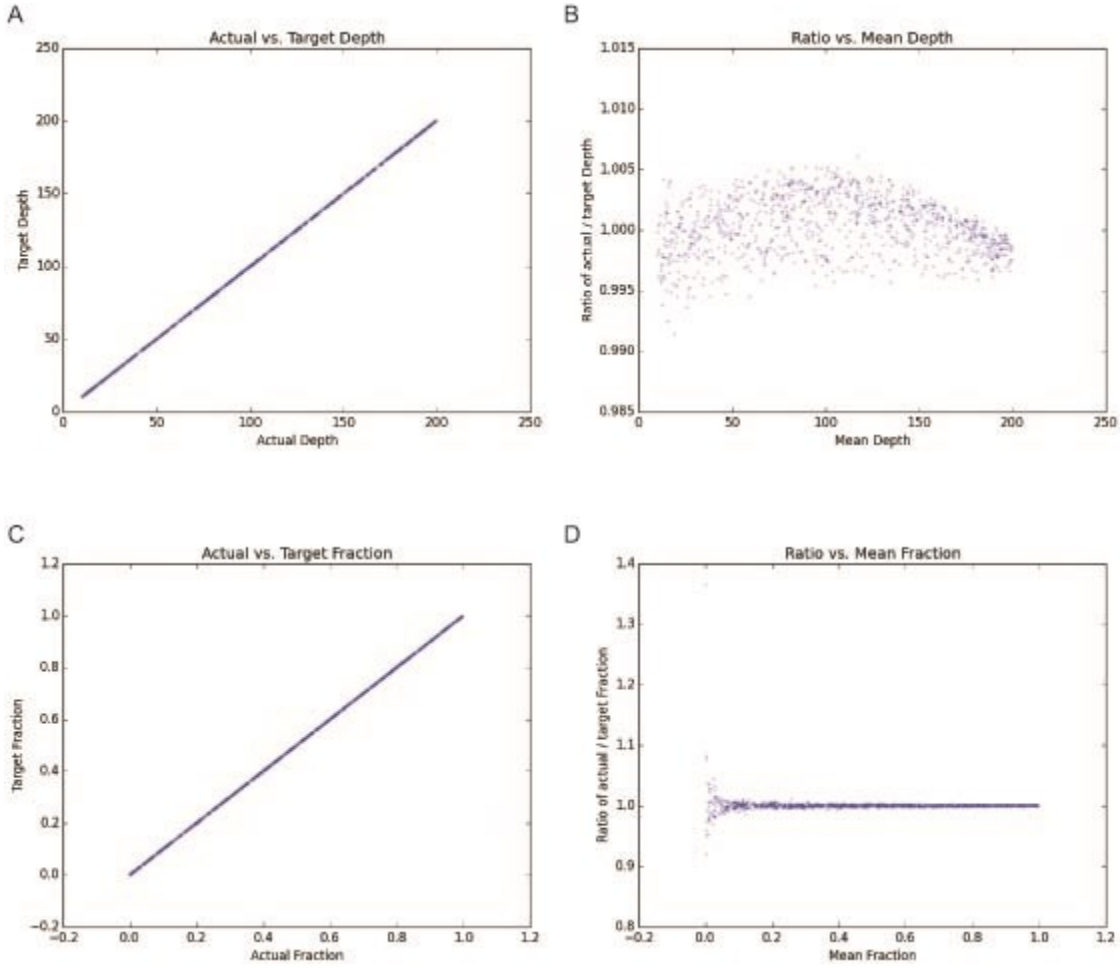
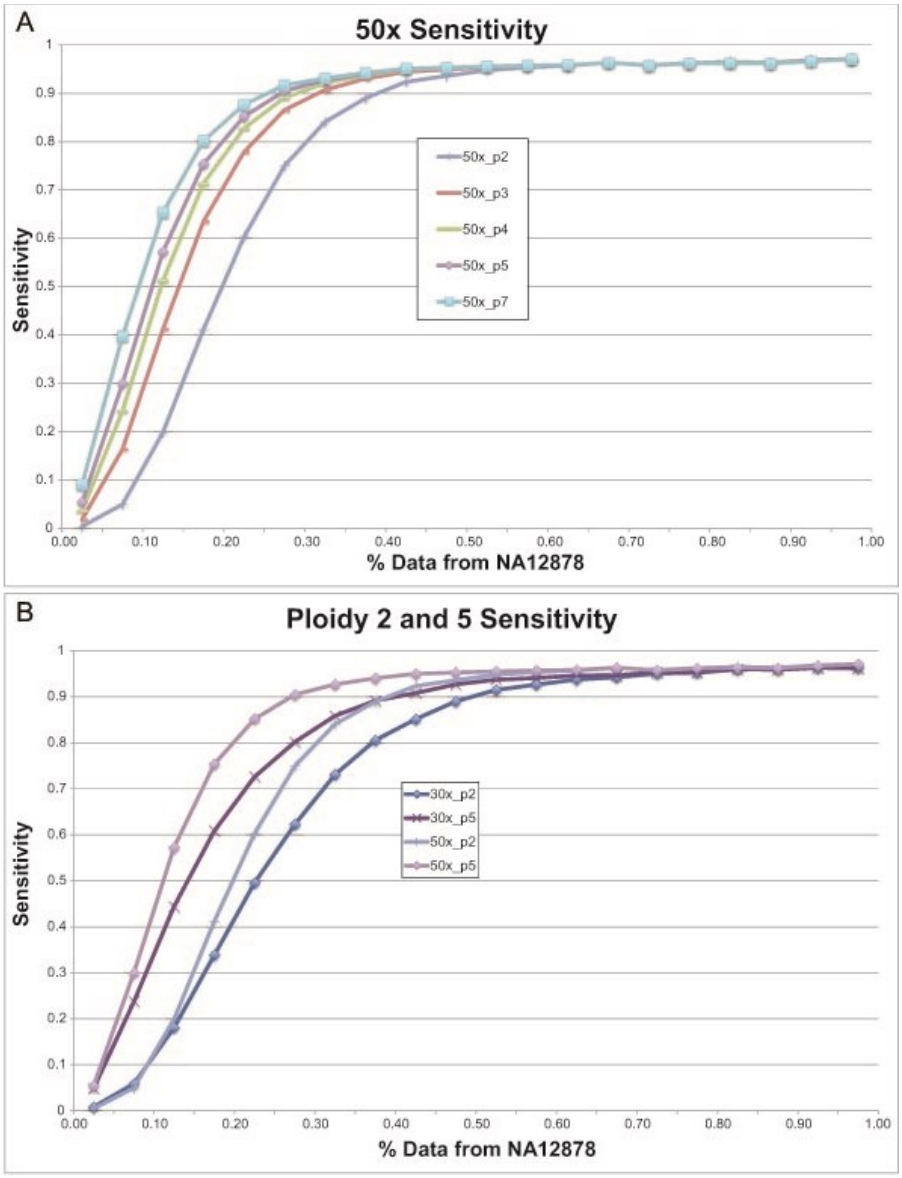


Figure 3.3. Sensitivity of the GATK HaplotypeCaller for mosaic variation.

(A) Sensitivity of the HaplotypeCaller for simulated mosaic variants from real sequence data at a sequencing read depth of 50x using various ploidy arguments. Higher ploidies resulted in higher sensitivity. (B) Comparison of ploidy 2 and ploidy 5 at 30x and 50x depth. Higher depths and high levels of the ploidy argument resulted in increased sensitivity.



3.3.3 Mosaic mutations in the Simons Simplex Collection

Given that mosaic variants may be identified with germline variant callers, we sought to determine the mosaic status of variants in the Simons Simplex Collection (SSC), a large collection of simplex autism pedigrees [249]. Extensive phenotypic data and whole-exome sequence data have been generated for all members of the collection and two non-overlapping callsets have been generated from the SSC exomes [212,228]. To increase sensitivity for detection of mosaic variants we performed a complete re-calling of all samples in the SSC with a –ploidy 5 setting (**Figure 3.4**). Variant filtration was performed using the GATK's variant quality score recalibration (VQSR) pipeline and *de novo* variants were identified using find_denovo with default parameters. This resulted in the identification of 6,408 *de novo* variants, of which 3,355 and 228 were present in the Iossifov or Krumm callsets, respectively and 2,825 were unique to our callset. Average coverage of high quality sequence reads at positions with identified *de novo* variants was 94.6, indicating that mosaic variants are likely to be accurately detected, when they occur. Of variants identified by Iossifov *et al.* or Krumm *et al.* but excluded from our callset the majority were filtered by VQSR (**Table 3.8**). We excluded fifteen families who had a child with more than 10 *de novo* mutations since these likely occur due to technical artifacts. After exclusion of these families, our callset contained 4,909 *de novo* variants. Variant effects were annotated using SnpEff and mosaic variants were identified from a binomial test with false-discovery protection using the Benjamini-Hochberg procedure [238]. This resulted in the identification of 1,036 mosaic variants at an FDR of 5%. Hereafter, we will refer to the variants identified as *de novo* but not mosaic as germline *de novo* variants.

Figure 3.4. Overview of the pipeline for calling variants in the Simons Simplex Collection.

For each step the total size of compressed data is given (e.g. 20 TB), number of individuals or files, analysis tool (e.g. Genome Analysis ToolKit [GATK] argument), average runtime job, and Amazon Web Services (AWS) EC2 instance type used in the analysis (e.g. c3.xlarge). Abbreviations: gVCF, genomic Variant Call Format file; glusterFS, a scalable network file system; NAS, network attached storage. Analyses were performed using AWS except the variant concatenation which was performed on a local cluster.

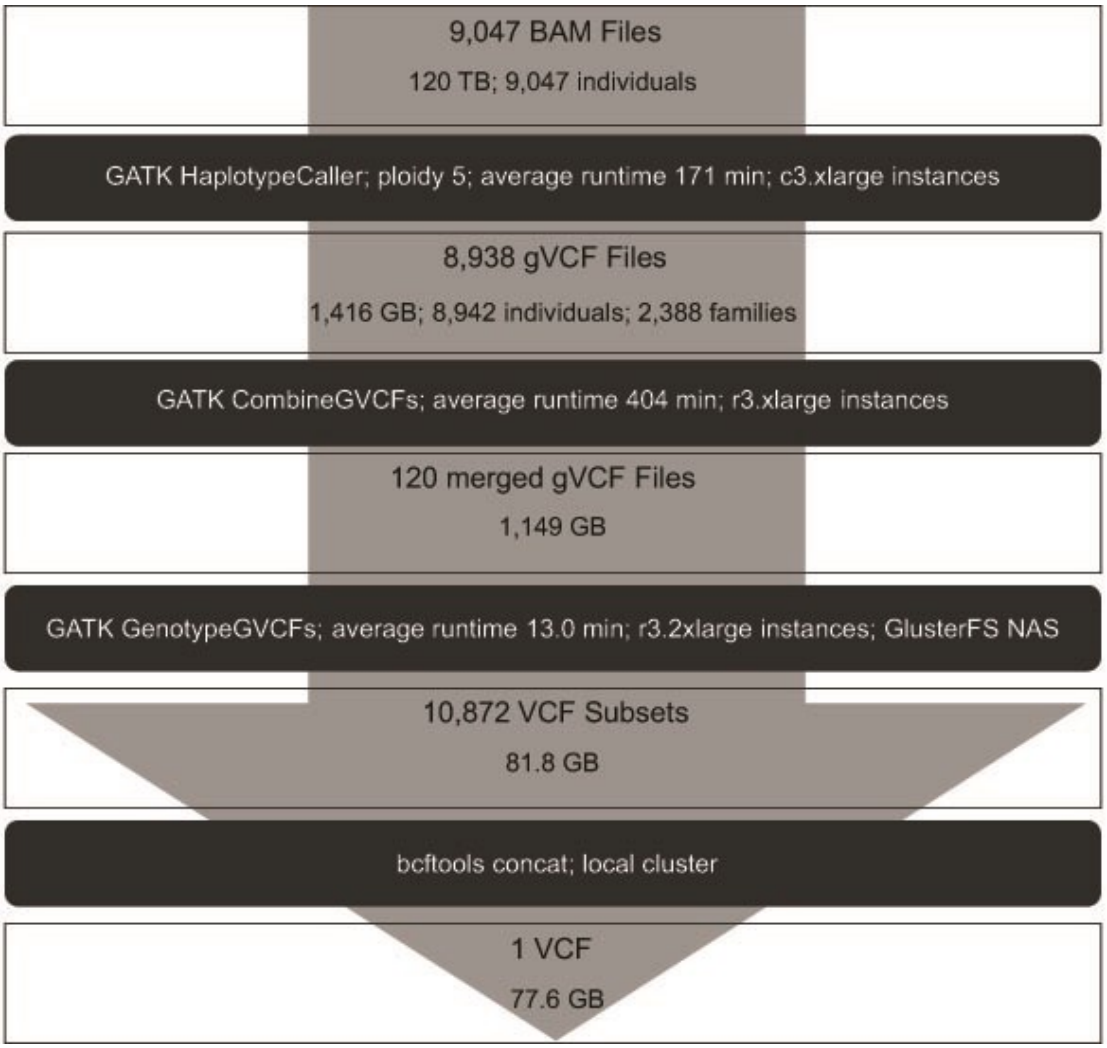


Table 3.8. Summary of the relationship between the Krumm/Iossifov callsets and the current callset.

Filters are applied from left to right. The VQSR column shows the number of variants filtered by the GATK's Variant Quality Score Recalibration pipeline. The columns called in father and called in mother indicate that the variant was genotyped as present in the father or mother, respectively. The column child alt DP indicates that fewer than three sequence reads passing quality filters were observed to support the alternate allele in the child. The columns child DP, father DP, and mother DP indicate a depth of less than 20 sequence reads were observed in the child, father, or mother, respectively. The columns child PL, father PL, and mother PL indicate a phred-scaled likelihood for the presence of the genotype (in the child) or the absence of the genotype (in the parents) of less than 20. The column multiple individuals shows the number of variants filtered due to their presence in more than two individuals.

	Earlier Filters <---------- > Later Filters																
	Starting total	Sample missing	Not called	VQSR	Called in father	Called in mother	Child alt DP	Child DP	Father DP	Mother DP	Child PL	Father PL	Mother PL	Multiple individuals	remaining	filtered
Krumm	1544	28	169	732	3	2	4	100	100	52	0	57	48	21	228	1316
Iossifov	5690	335	141	995	21	63	4	89	35	39	0	98	100	415	3355	2335

To ensure that the variants in our callset were present in the samples, we performed Sanger sequencing of 97 (47 mosaic and 50 germline *de novo*) variants (**Table 3.9**, pre-filter). For all of our validation methods, we present both "detection precision" and "classification precision", where applicable. We define "detection precision" as our precision for the presence of the variant in the sample, while we define "classification precision" as our precision for the presence of the variant in the sample and correct classification of the variant as either mosaic or germline *de novo*. Of the 97 reactions, sequencing was informative for 76. The variant of interest was identified in 100% of samples when the variant was annotated as germline *de novo*. However, in samples harboring a mosaic variant, precision for the presence of the variant was modest (54%). Mosaic variants that failed validation were often called with few reads supporting the alternate allele and were frequently called uniquely in our callset. To improve downstream analyses, we made the conservative choice of requiring that identified mosaic variants be present jointly in our callset and in the Iossifov or Krumm callsets. This filter greatly improves the precision of our callset (100% for variant presence; **Table 3.9**, post-filter) with little change in sensitivity. However, the precision of the classification of the mosaic status of variants remained modest (68%).

Table 3.9. Sanger sequencing validation of variants in the Simons Simplex Collection.

Pre-filter mosaic variants were identified as described in the methods section. Post-filter mosaic variants have the additional requirement that they must be identified jointly in the current callset and one of the callsets produced by Iossifov *et al.* or Krumm *et al.* Assay success refers to technical success of the sequencing assay.

		Chosen	Assay Success	Variant Present	Detection Precision	Variant Mosaic	Classification Precision
Pre-filter	Mosaic	47	37	20	0.54	14	0.37
	Germline *de novo*	50	39	39	1.00	3	0.92
Post-filter	Mosaic	26	19	19	1.00	13	0.68
	Germline *de novo*	50	39	39	1.00	3	0.92

While Sanger sequencing provides an accurate assessment of the presence of *de novo* variants, examination of the chromatograms can provide only approximate estimation of mosaic status. To more accurately assess the mosaic status of the identified variants, we performed sequence read phasing of all identified *de novo* variants. Phasing of the potential mosaic variants relative to nearby inherited heterozygous variants using sequence reads may rigorously confirm the presence of mosaicism. This occurs when three parental haplotypes are inferred: a single haplotype from one parent (e.g. having the minor allele of the neighboring SNP), and two haplotypes from the other parent (e.g. having the major allele of the neighboring SNP) which resolve into a haplotype with the mosaic allele and a distinct haplotype lacking the mosaic allele (**Figure 3.5**). We wrote a program called phase-mosaic to perform phasing validation (see Materials and Methods). Of the variants passing filters, phasing was informative for 51 mosaic variants of which 29 were validated as mosaic (57%; **Table 3.10**, Pre-filter). Mosaic variants identified by next-generation sequencing that failed phasing confirmation tended to be variants called at high depth with a high AARF. We suspect that these variants appear to be mosaic due to preferential capture of the reference allele during exome enrichment. To correct for this effect, we modified our criteria for the identification of mosaic variants to require that mosaic variants have AARF less than 34%. With these adjusted parameters, the precision of our classification of mosaic status improved to 87% (**Table 3.10**, Post-filter).

Figure 3.5. Confirmation of mosaic status by sequence read phasing.

Phasing of a *de novo* variant confirms mosaic status of a variant in *OR4M2* in individual 12977.s1. The heterozygous variant on the left is inherited (green pattern; unknown parental origin). The deletion on the right was identified as occurring *de novo*. The presence of multiple reads (box) containing the inherited allele but not the *de novo* variant demonstrates the occurrence of three distinct haplotypes implying post-zygotic origin and mosaicism.

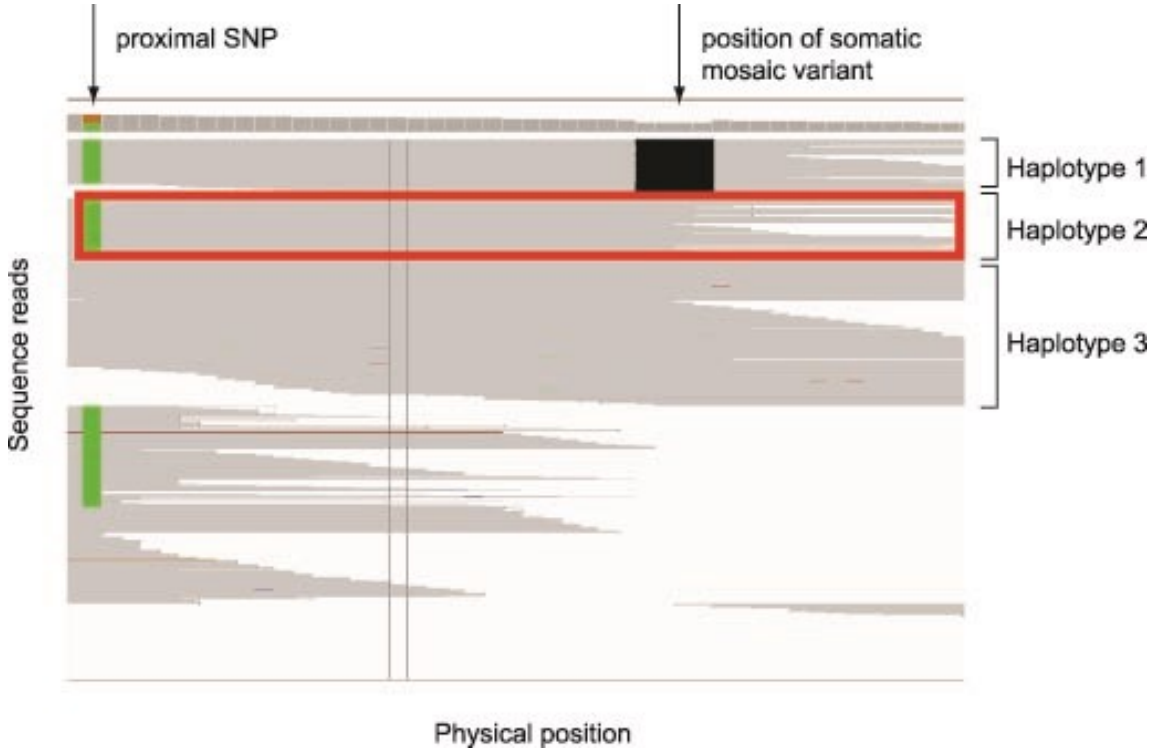


Table 3.10. Read-backed phasing validation of the mosaic status of identified variants.

Pre-filter mosaic variants were identified using the binomial test resulting in moderate sensitivity for mosaic status. Post-filter mosaic variants had the added requirement of an AARF of less than 34%, improving classification precision.

		Variants	Phasing Informative	Mosaic	Germline *de novo*	Classification Precision
Pre-filter	Mosaic	342	51	29	22	0.57
	Germline *de novo*	3742	443	27	416	0.94
Post-filter	Mosaic	221	30	26	4	0.87
	Germline *de novo*	3874	468	31	437	0.93

Validation of mosaic variants was also performed using pyrosequencing (**Table 3.11**). Likely-gene disrupting (LGD) and missense variants in probands across a range of allele frequencies were chosen for pyrosequencing validation (see Materials and Methods). Consistent with the post-filter results of Sanger sequencing and physical phasing, pyrosequencing validation demonstrated high precision for variant detection and variant classification (**Table 3.12**). Of all variants validated by an orthogonal sequencing technology, 16 were validated with multiple validation methods. The results were consistent, except for apparent inaccuracy in the classification of mosaic status by Sanger sequencing.

Table 3.11. Pyrosequencing validation of *de novo* variants in the Simons Simplex Collection.

Experiment date	Chromosome	Position	Family	Individual	Gene	Ref allele	Alt allele	% ref	% alt	p-mosaic	Notes
8/11/2015	13	51948834	14687	pM	INTS6	G	A	61	39	0.000261	
8/11/2015	11	93463107	14069	pM	CEP295	C	CA	46	54	0.010037	
8/11/2015	3	20161089	11592	pM	KAT2B	G	A	49	51	9.76E-06	
8/11/2015	13	92345513	11042	pM	GPC5	T	C	69	31	2.45E-07	
8/11/2015	1	3385486	13350	pM	ARHGEF16	C	T	79	21	0.001444	
8/7/2015	19	55439074	13199	pF	NLRP7	C	G	79	21	1.79E-05	
8/7/2015	15	66198459	12056	pM	MEGF11	G	A	69	31	8.69E-12	
8/7/2015	13	51948834	14687	pM	INTS6	G	A	65	35	0.000261	Warning
8/7/2015	13	92345513	11042	pM	GPC5	T	C	64	36	2.45E-07	
8/7/2015	12	116424952	14416	pM	MED13L	C	T	46	54	0.586798	
8/7/2015	11	64939990	14592	pM	SPDYC	C	T	46	54	0.623983	
8/7/2015	9	95277357	13364	pM	ECM2	G	A	49	51	0.872848	Warning
8/7/2015	2	234676519	14644	pM	UGT1A10	C	T	53	47	0.918288	
8/7/2015	2	127944869	14307	pM	CYP27C1	AC	A	53	47	0.924205	
8/7/2015	7	104748100	12952	pM	KMT2E	TC	T	35	65	0.99024	
8/7/2015	15	41192955	13575	pM	VPS18	C	T	75	25	0.006765	
8/7/2015	2	6990025	13254	pM	CMPK2	G	A	50	50	0.032623	
8/7/2015	13	21205235	14226	pM	IFT88	T	G	53	47	0.040345	
8/7/2015	8	124787443	12221	pM	FAM91A1	C	T	50	50	0.015241	
8/3/2015	7	100850973	13846	pM	PLOD3	C	T	70	30	9.85E-07	
7/28/2015	7	100850973	13846	pM	PLOD3	C	T	75	25	9.85E-07	
7/17/2015	10	876865	14093	pM	LARP4B	TCA	T	90	10	3.62E-14	
7/17/2015	4	80954694	13848	pM	ANTXR2	C	T	65	35	5.24E-09	
7/2/2015	10	876865	14093	pM	LARP4B	TCA	T	81	19	3.62E-14	
7/2/2015	11	9111272	13506	pM	SCUBE2	C	A	63	37	0.012983	Warning

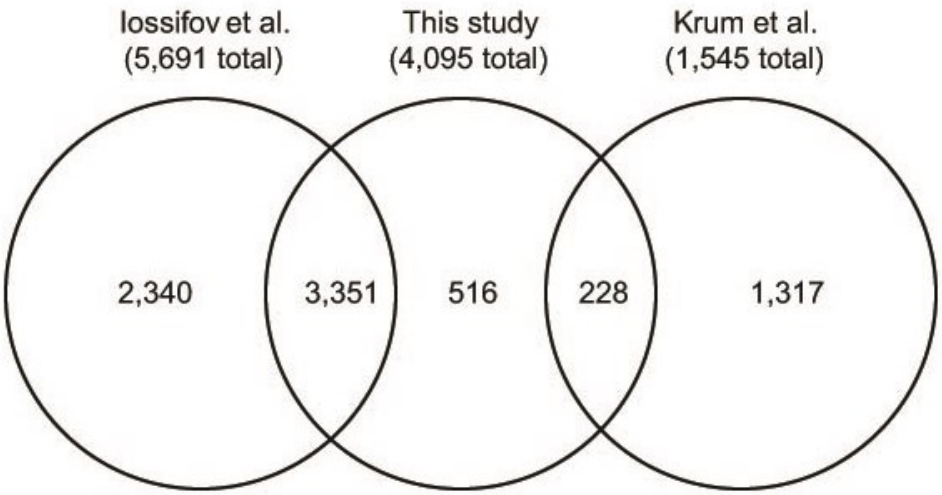
Experiment date	Chromosome	Position	Family	Individual	Gene	Ref allele	Alt allele	% ref	% alt	p-mosaic	Notes
7/2/2015	2	234676519	14644	pM	UGT1A10	C	T	43	57	0.918288	

Table 3.12. Pyrosequencing validation of variants in the Simons Simplex Collection.

	Successful validation	Variant present	Detection precision	Variant mosaic	Classification precision
Mosaic	11	11	1.00	9	0.82
Germline *de novo*	10	10	1.00	1	0.90

After the application of filters, we identified a total of 4,095 *de novo* variants in our high-confidence callset, 221 of which were classified as mosaic. Based on our validation experiments, we estimate that our precision for the presence of the called variants is near 100% with the precision of the classification of mosaic variants measured at 87% or 82% by phasing or pyrosequencing, respectively. Of the variants in our final callset, 3,351 appear jointly in the current callset and the callset produced by Iossifov *et al.* while 228 appear jointly in the current callset and the callset produced by Krumm *et al.* (**Figure 3.6**). In the high confidence callset no mosaic mutations were identified that were shared between a sibling pair.

Figure 3.6. Venn diagram of variants in the high-confidence callset.



3.3.4 Properties of mosaic variants

To better understand the properties of variants in our callset, we examined the mutational spectra of the identified mosaic variants relative to germline *de novo* variants (**Table 3.13**). We find that mosaic variants have significantly more deletions than germline *de novo* variants (Fisher's exact test, p = 5.2e-4). However, the rate of occurrence of other types of mosaic mutations is approximately equal to the rate of occurrence of the corresponding *de novo* mutation. The relative enrichment of mosaic mutations for deletions may indicate an increased rate of false-positive mutation as the identification of indels from next-generation sequence data is known to be difficult. However, our precision when validating mosaic mutation was quite high (see above). We attempted validation of four deletions in the high-confidence callset using Sanger sequencing. All four deletions were found in the sample and three of four were confirmed as mosaic. Besides false-positives, the enrichment of mosaic mutations may indicate a non-reference allele bias, where germline *de novo* deletions are occurring in the samples but are incorrectly classified as mosaic due to mapping errors. Phasing assessed the mosaic status of eight mosaic deletions, six of which were confirmed as mosaic resulting in a classification precision of 75%, slightly less than the overall classification precision of 87% from sequence read phasing. Therefore, inaccurate classification of the mosaic status of *de novo* deletions may contribute to the observed enrichment. An additional hypothesis is that the mechanism underlying mosaic mutation wholly or partly differs from that of germline *de novo* mutation and the relative enrichment of deletions may be attributed to these differences in underlying mechanisms.

Table 3.13. The mutation spectra of mosaic variants relative to germline *de novo* variants.

Odds ratios of less than 1 indicate mosaic variants are relatively depleted for those events while odds ratios of greater than 1 indicate relative enrichment. p values were calculated using a Fisher's exact test.

Mutation Type	Odds Ratio	p-value
transition	0.88	0.265
transversion	1.10	0.503
insertion	0.83	1.000
deletion	2.49	0.001
CpG	1.10	0.506

3.3.5 Rates of mosaic mutation

Previous studies have indicated that *de novo* mutations occur at higher rates in probands relative to controls leading to the implication of *de novo* variants as contributing to disease diagnoses [212,230]. We utilized our high-confidence mosaic variant callset to compare the rates of mosaic and germline *de novo* mutation in probands relative to unaffected siblings. Following the protocol of Iossifov *et al.* we defined regions of joint 40x coverage in children of quad families and extrapolated rates of mutation within these joint 40x regions to the entire capture region (**Figure 3.7**; **Table 3.14**). Consistent with previous results, we find that germline *de novo* LGD mutations are significantly enriched in probands relative to controls (p=0.001). In addition we find that all classes of mosaic mutations are significantly enriched in probands (p=0.003). Interestingly, we observe contribution to disease from all classes of mosaic variation, whereas the contribution of germline *de novo* variation to disease is primarily from LGD mutations.

Fig 3.7. Rates of mutation in the Simons Simplex Collection.

Average number of mutations per exome, as calculated using joint 40x regions. Error bars represent the 95% confidence interval for the mean. (A) Mutations categorized as mosaic with AARF < 0.34 and q < 0.05. (B) Germline *de novo* mutations as determined by q > 0.05 or AARF > 0.34.

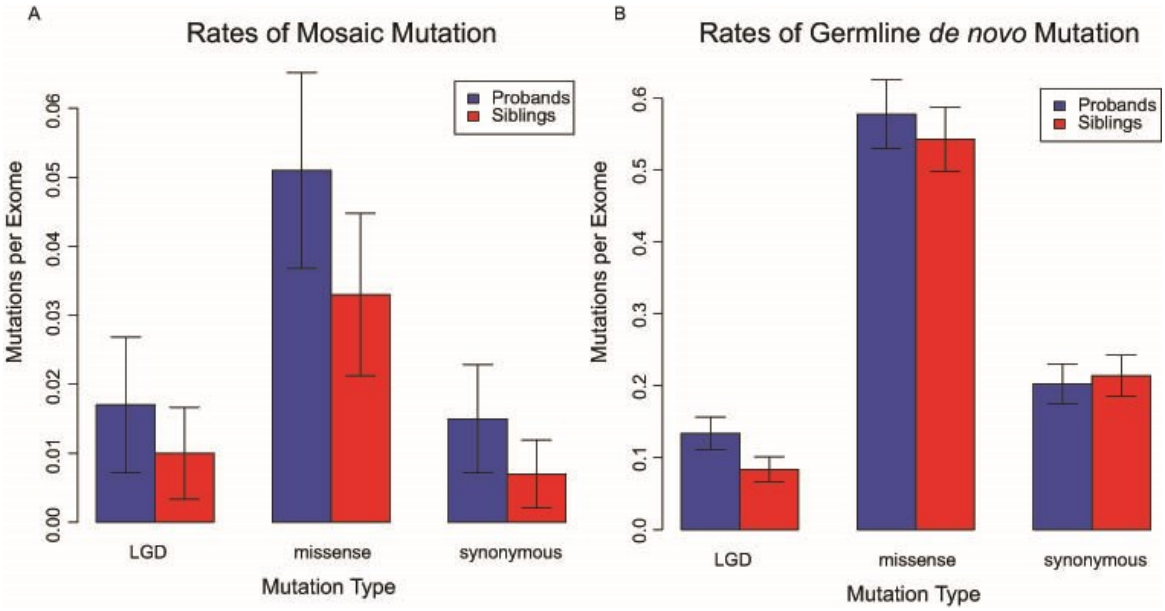


Table 3.14. Rates of *de novo* mutation in individuals in the Simons Simplex Collection.

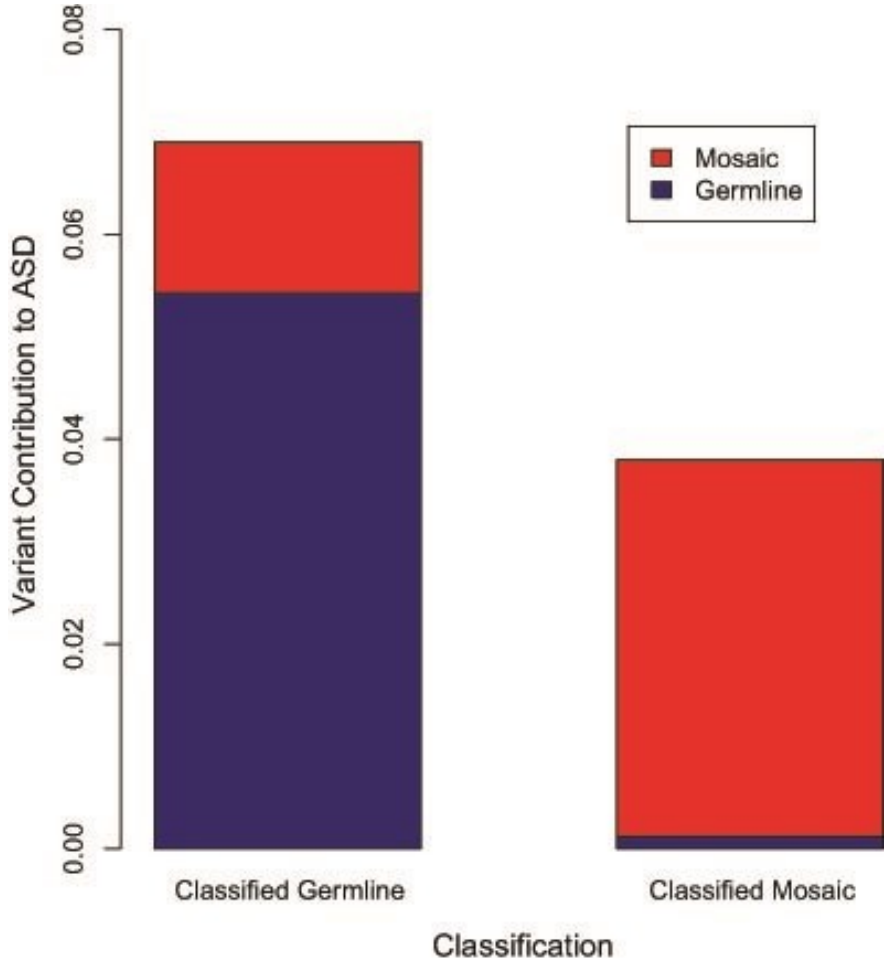
Rates of mutation were measured using joint 40x regions in probands and siblings in quad families as described in the Methods section.

	Probands mean rate	Probands sd	Siblings mean rate	Siblings sd	Fraction contributing	Contributes to	Proband mean 95% CI	Sibling mean 95% CI	p-value
non_mosaic_lgd	0.134	0.486	0.084	0.374	0.373	0.050	0.02272	0.01748	0.001
non_mosaic_missense	0.578	1.022	0.543	0.949	0.060	0.034	0.04775	0.04437	0.301
non_mosaic_synonymous	0.203	0.595	0.214	0.616	-0.058	-0.012	0.02780	0.02877	0.563
non_mosaic_noncoding	0.011	0.133	0.016	0.163	-0.396	-0.004	0.00619	0.00760	0.379
non_mosaic_other	0.057	0.333	0.056	0.329	0.021	0.001	0.01555	0.01539	0.914
non_mosaic_all	0.982	1.335	0.913	1.254	0.071	0.069	0.06241	0.05863	0.113
mosaic_lgd	0.017	0.211	0.010	0.143	0.395	0.007	0.00985	0.00666	0.263
mosaic_missense	0.051	0.303	0.033	0.252	0.344	0.018	0.01418	0.01178	0.063
mosaic_synonymous	0.015	0.168	0.007	0.105	0.534	0.008	0.00786	0.00492	0.095
mosaic_noncoding	0.000	0.000	0.000	0.000	-	0.000	0.00000	0.00000	-
mosaic_other	0.007	0.120	0.001	0.044	0.847	0.006	0.00562	0.00206	0.046
mosiac_all	0.090	0.442	0.052	0.309	0.424	0.038	0.02067	0.01444	0.003
n-probands	1758								
n-sibs	1758								

To account for errors in the classification of variants as either mosaic or germline *de novo*, we extended our model of contributory variation to include incorrectly classified variants. In this model, mosaic variants incorrectly classified as germline *de novo* account for a substantial portion of the genetic contribution of variants classified as germline *de novo* (**Figure 3.8**). In total, mosaic variation contributes to 5.1% of ASD cases (95% credible interval [CI], 1.3% to 8.9%) while all classes of germline *de novo* variation contribute to 5.6% of ASD cases (95% CI, 1.8% to 9.4%). The percent of contributory variants to total variants are measured as 6.0% (95% CI, 2.0% to 10%) and 33% (95% CI, 9.6% to 54%) for germline *de novo* and mosaic variants, respectively.

Figure 3.8. The contribution of *de novo* mutations to ASD.

The contribution of classified mosaic and germline mutations are shown. For each classification, the contribution is divided into correctly classified and incorrectly classified variation. The contribution of incorrectly classified mosaic variants (called germline; left bar, upper region) is substantial, but the contribution of incorrectly classified germline variation (called mosaic; right bar, lower region) is small.



3.3.6 Functional consequences of de novo mutation in the SSC

While differences in the rates of mutation in affected individuals may implicate mutations in disease, it may also be the case that mutations in probands occur in more functionally conserved genomic regions. Using all of the mutations in our high-confidence callset, we test the hypothesis that mutations in probands occur at more conserved genomic regions. For this analysis, we use three measures of conservation: base-level conservation as measured by PhyloP, and gene-level conservation as measured by HomoloGene or ExAC (**Table 3.15**) [240-242]. The gene-level conservation measures from ExAC and Homologene are complimentary as HomoloGene provides a measure of evolutionary conservation while ExAC provides a measure of conservation in extant human populations. We find that germline *de novo* LGD mutations occur at more highly conserved positions in probands relative to controls as measured by PhyloP score ($p = 0.048$, effect = 0.54). For mosaic missense mutations, we observe a stronger effect (0.91), although the test does not reach statistical significance due to the small sample size ($p = 0.179$). We find that germline *de novo* missense variants occur significantly more often in genes thought to be intolerant of loss-of-function mutation as annotated by ExAC ($p = 0.013$). While our analysis does not show that germline *de novo* missense mutations occur at significantly higher rates in probands relative to siblings ($p = 0.30$), germline *de novo* missense variants likely target genes less tolerant of functional mutation more frequently in affected individuals relative to their siblings.

Table 3.15. Conservation at sites of *de novo* variation in probands and siblings.

Each variant was annotated with measures of conservation as described in the Methods section. Measures of conservation were compared between probands and unaffected siblings and p values were calculated using a Wilcoxon rank sum test. Estimates of effect size were calculated as the estimated difference in ranks.

	PhyloP Score		HomoloGene		ExAC	
	Effect_Size	p-value	Effect_Size	p-value	Effect_Size	p-value
mosaic_missense	0.91	0.179	6.90E-05	0.222	2.29E-05	0.534
non-mosaic_missense	0.16	0.250	-6.33E-05	0.735	-1.02E-06	0.013
mosaic_LGD	0.46	0.679	4.70E-05	0.808	3.27E-05	0.894
non-mosaic_LGD	0.54	0.049	-3.95E-05	0.876	-2.10E-05	0.072

The initial publication of the SSC exome sequencing data demonstrated enrichment of mutations in specific classes of gene targets [212]. To find insight into the mutational mechanisms and functional consequences of mosaic mutation, we replicated this analysis (with modification) using our high-confidence callset. This analysis confirmed the significant enrichment of germline *de novo* LGD mutations from probands in FMRP targets, chromatin modifiers, and genes with known LGD mutations in intellectual disability or schizophrenia (**Table 3.16**). In addition, we observed enrichment of mosaic missense and LGD mutations in probands and siblings in genes involved in embryonic development (18 observed versus 12.6 expected for probands; 9 observed versus 6.7 expected for siblings), however this enrichment did not reach statistical significance ($p = 0.12$ for probands; $p = 0.30$ for siblings). We also tested for overlap between genes targeted by mosaic missense and LGD mutations and a set of 107 genes that had been strongly implicated in ASD using the null-length model (see Methods) [210]. We found three of the 98 genes with mosaic missense or LGD mutations in probands have been previously implicated in ASD (*KMT2C*, *NCKAP1*, and *MYH10*). The presence of the *NCKAP1* was confirmed by Sanger sequencing, but did not confirm its mosaic status. However, the number of mosaic mutations in ASD genes does not reach statistical significance for enrichment in the set of 107 ASD genes (3 observed; 1.15 expected; $p = 0.109$). Zero of 52 genes targeted by mosaic missense or mosaic LGD mutations in siblings were previously implicated in ASD.

Table 3.16. Gene target overlap.

Columns describe gene target enrichment analysis.

class	Set FMRP Targets	Set PSD	Set Embryonic	Set Chromatin Modifiers	Set Essential Genes	Set Mendelian Disease Genes	Set De Novo LGDs In Schizoprenia	Set De Novo LGDs In ID
Dnv mos functional prb	6 (9.46, 0.30384)	10 (9.10, 0.72682)	18 (12.57, 0.12817)	3 (3.29, 1.00000)	15 (11.92, 0.35185)	3 (2.24, 0.49327)	1 (1.03, 1.00000)	0 (0.33, 1.00000)
Dnv mos functional sib	5 (5.02, 1.00000)	8 (4.83, 0.14541)	9 (6.67, 0.30267)	2 (1.74, 0.69389)	4 (6.32, 0.40126)	1 (1.19, 1.00000)	0 (0.55, 1.00000)	0 (0.17, 1.00000)
Dnv nm LGD prb	30 (18.53, 0.00957)	23 (17.82, 0.21182)	32 (24.62, 0.12944)	14 (6.44, 0.00712)	32 (23.35, 0.06041)	4 (4.40, 1.00000)	6 (2.01, 0.01647)	5 (0.64, 0.00051)
Dnv nm LGD sib	6 (9.56, 0.30490)	13 (9.19, 0.22135)	11 (12.70, 0.76324)	3 (3.32, 1.00000)	8 (12.04, 0.28002)	2 (2.27, 1.00000)	1 (1.04, 1.00000)	1 (0.33, 0.28165)

3.4 Discussion

There are three major conclusions from this study. First, we show that mosaic mutations occur frequently in individuals diagnosed with ASD and their unaffected siblings. We identify a total of 4,095 *de novo* mutations, of which 221 (5.4%) are classified as mosaic. This is similar to previously reported estimates for the fraction of mosaic variants in individuals with intellectual disability [233]. In light of previous work demonstrating the presence of mosaic mutation in diverse body tissues we believe that the mosaic mutations we identified are not unique to blood but are dispersed throughout the body [202]. Although the early steps of our pipeline were performed explicitly to increase our sensitivity for mosaic mutation, many of our filtering steps were conservative and we likely underestimate the true fraction of mosaic mutations in the Simons Simplex Collection. Our filtering approach combined with recent improvements in variant detection algorithms likely accounts for most of the differences between our variant callset and the callsets published by Iossifov *et al.* and Krumm *et al.*

Second, we find that mosaic mutations are significantly enriched in probands relative to their siblings. Using our model of contributory variation we estimate that 33% of mosaic mutations contribute to 5.1% of ASD diagnoses. As mosaic mutations arise post-zygotically in only a fraction of the cells of an individual, we expect that these results have implications for the interpretation of twin studies, especially observed cases of phenotypic discordance between monozygotic twins.

Third, we find that tissue-specific mosaic mutations do not occur in the paired samples at our limit of detection. Given the lack of publications on validated tissue-specific mutations in tissues without visual abnormality and the absence of brain-specific mutation in

a centenarian [246], we do not believe that this finding is unexpected. While a recent study reported tissue-specific mosaic mutation [250], the results presented here include validation of detected mutations showing that, in our study, these were false positive findings. It is possible that tissue-specific mutations do contribute to ASD in at least some cases. However, discovery of such variation and its implication in disease may require larger numbers of samples or more sensitive approaches (such as single-cell sequencing).

Together, these results indicate that mosaic mutations are an identifiable subset of *de novo* mutation. As heritable factors that may arise in a single twin of a monozygotic pair [231,232], contributory mosaic mutation implies some expected level of discordance between monozygotic twins due to heritable factors arising post-zygotically. Furthermore, high-confidence identification of contributory mosaic mutation in affected probands implies a lower risk of familial recurrence in some families.

References

1. Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res 17: 6463-6471.
2. Barrett AJ, Battiwalla M (2010) Relapse after allogeneic stem cell transplantation. Expert review of hematology 3: 429-441.
3. Yang DY, Watt K (2005) Contamination controls when preparing archaeological remains for ancient DNA analysis. Journal of Archaeological Science 32: 331-336.
4. Edwards JH (1989) Familiarity, recessivity and germline mosaicism. Ann Hum Genet 53: 33-47.
5. Hartl DL (1971) Recurrence risks for germinal mosaics. Am J Hum Genet 23: 124-134.
6. Poduri A, Evrony GD, Xuyu C, Walsh CA (2013) Somatic Mutation, Genomic Variation, and Neurological Disease. Science 341.
7. Campbell IM, Yuan B, Robberecht C, Pfundt R, Szafranski P, et al. (2014) Parental Somatic Mosaicism Is Underrecognized and Influences Recurrence Risk of Genomic Disorders. Am J Hum Genet 95: 173-182.
8. van der Maarel SM, Deidda G, Lemmers RJLF, van Overveld PGM, van der Wielen M, et al. De Novo Facioscapulohumeral Muscular Dystrophy: Frequent Somatic Mosaicism, Sex-Dependent Phenotype, and the Role of Mitotic Transchromosomal Repeat Interaction between Chromosomes 4 and 10. The American Journal of Human Genetics 66: 26-35.
9. Boveri T (1929) The Origin of Malignant Tumors. Boveri M, translator. Baltimore, Maryland: The Williams and Wilkins Company. 119 p.
10. Knudson AG (1971) Mutation and Cancer: Statistical Study of Retinoblastoma. Proceedings of the National Academy of Sciences 68: 820-823.
11. Nowell PC (1976) The clonal evolution of tumor cell populations. Science 194: 23-28.
12. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., et al. (2013) Cancer genome landscapes. Science 339: 1546-1558.
13. McClintock B (1951) Chromosome organization and genic expression. Cold Spring Harb Symp Quant Biol 16: 13-47.
14. Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J (2009) Massively parallel exon capture and library-free resequencing across 16 genomes. Nat Meth 6: 315-316.
15. Brack C, Hirama M, Lenhard-Schuller R, Tonegawa S (1978) A Complete Immunoglobulin Gene is Created by Somatic Recombination. Cell 15: 1-14.
16. Tonegawa S (1983) Somatic generation of antibody diversity. Nature 302: 575-581.
17. Krangel MS (2009) Mechanics of T cell receptor gene rearrangement. Curr Opin Immunol 21: 133-139.
18. Frumkin D, Wasserstrom A, Kaplan S, Feige U, Shapiro E (2005) Genomic variability within an organism exposes its cell lineage tree. PLoS Comput Biol 1: e50.
19. Hoeijmakers JHJ (2009) DNA Damage, Aging, and Cancer. New England Journal of Medicine 361: 1475-1485.
20. Kennedy SR, Loeb LA, Herr AJ (2012) Somatic mutations in aging, cancer and neurodegeneration. Mechanisms of Ageing and Development 133: 118-126.
21. Jeppesen DK, Bohr VA, Stevnsner T (2011) DNA repair deficiency in neurodegeneration. Progress in Neurobiology 94: 166-200.

22. Erickson RP (2003) Somatic gene mutation and human disese other than cancer. Mutat Res 543: 125-136.
23. Erickson RP (2010) Somatic gene mutation and human disease other than cancer: an update. Mutat Res 705: 96-106.
24. Erickson RP (2014) Recent advances in the study of somatic mosaicism and diseases other than cancer. Current Opinion in Genetics & Development 26: 73-78.
25. Hirschhorn R (2003) In vivo reversion to normal of inherited mutations in humans. J Med Genet 40: 721-728.
26. Jonkman MF, Castellanos Nuijts M, van Essen AJ (2003) Natural repair mechanisms in correcting pathogenic mutations in inherited skin disorders. Clin Exp Dermatol 28: 625-631.
27. Lai-Cheong JE, McGrath JA, Uitto J (2011) Revertant mosaicism in skin: natural gene therapy. Trends Mol Med 17: 140-148.
28. Jonkman MF (1999) Revertant mosaicism in human genetic disorders. Am J Med Genet 85: 361-364.
29. Happle R (1987) Lethal genes surviving by mosaicism: a possible explanation for sporadic birth defects involving the skin. J Am Acad Dermatol 16: 899-906.
30. Liu P, Carvalho CM, Hastings PJ, Lupski JR (2012) Mechanisms for recurrent and complex human genomic rearrangements. Curr Opin Genet Dev 22: 211-220.
31. Cimini D, Howell B, Maddox P, Khodjakov A, Degrassi F, et al. (2001) Merotelic Kinetochore Orientation Is a Major Mechanism of Aneuploidy in Mitotic Mammalian Tissue Cells. The Journal of Cell Biology 153: 517-528.
32. Robinson WP (2000) Mechanisms leading to uniparental disomy and their clinical consequences. Bioessays 22: 452-459.
33. Kotzot D (2008) Complex and segmental uniparental disomy updated. J Med Genet 45: 545-556.
34. Conlin LK, Thiel BD, Bonnemann CG, Medne L, Ernst LM, et al. (2010) Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. Human Molecular Genetics 19: 1263-1275.
35. Liehr T (2010) Cytogenetic contribution to uniparental disomy (UPD). Mol Cytogenet 3: 8.
36. Hancks DC, Kazazian HH, Jr. (2012) Active human retrotransposons: variation and disease. Curr Opin Genet Dev 22: 191-203.
37. van den Hurk JA, Meij IC, Seleme MC, Kano H, Nikopoulos K, et al. (2007) L1 retrotransposition can occur early in human embryonic development. Hum Mol Genet 16: 1587-1592.
38. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, et al. (2012) Landscape of Somatic Retrotransposition in Human Cancers. Science 337: 967-971.
39. Kim JC, Mirkin SM (2013) The balancing act of DNA repeat expansions. Curr Opin Genet Dev 23: 280-288.
40. Mirkin SM (2007) Expandable DNA repeats and human disease. Nature 447: 932-940.
41. Ueno S-i, Kondoh K, Komure Y, Komure O, Kuno S, et al. (1995) Somatic mosaicism of CAG repeat in dentatorubral-pallidoluysian atrophy (DRPLA). Human Molecular Genetics 4: 663-666.

42. Hashida H, Goto J, Suzuki T, Jeong S-Y, Masuda N, et al. (2001) Single cell analysis of CAG repeat in brains of dentatorubral-pallidoluysian atrophy (DRPLA). Journal of the Neurological Sciences 190: 87-93.
43. Hellenbroich Y, Schwinger E, Zühlke CH (2001) Limited somatic mosaicism for Friedreich's ataxia GAA triplet repeat expansions identified by small pool PCR in blood leukocytes. Acta Neurologica Scandinavica 103: 188-192.
44. Kahlem P, Djian P (2000) The expanded CAG repeat associated with juvenile Huntington disease shows a common origin of most or all neurons and glia in human cerebrum. Neuroscience Letters 286: 203-207.
45. McMurray CT (2010) Mechanisms of trinucleotide repeat instability during human development. Nat Rev Genet 11: 786-799.
46. Møllersen L, Rowe AD, Larsen E, Rognes T, Klungland A (2010) Continuous and Periodic Expansion of CAG Repeats in Huntington's Disease R6/1 Mice. PLoS Genet 6: e1001242.
47. Montermini L, Kish SJ, Jiralerspong S, Lamarche JB, Pandolfo M (1997) Somatic mosaicism for Friedreich's ataxia GAA triplet repeat expansions in the central nervous system. Neurology 49: 606-610.
48. Lindahl T, Wood RD (1999) Quality Control by DNA Repair. Science 286: 1897-1905.
49. Green P, Ewing B, Miller W, Thomas PJ, Green ED (2003) Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet 33: 514-517.
50. Polak P, Arndt PF (2008) Transcription induces strand-specific mutations at the 5' end of human genes. Genome Res 18: 1216-1223.
51. Gilbert DM (2001) Making Sense of Eukaryotic DNA Replication Origins. Science 294: 96-100.
52. Gilbert DM (2010) Evaluating genome-scale approaches to eukaryotic DNA replication. Nat Rev Genet 11: 673-684.
53. Levy MZ, Allsopp RC, Futcher AB, Greider CW, Harley CB (1992) Telomere end-replication problem and cell aging. J Mol Biol 225: 951-960.
54. van Echten-Arends J, Mastenbroek S, Sikkema-Raddatz B, Korevaar JC, Heineman MJ, et al. (2011) Chromosomal mosaicism in human preimplantation embryos: a systematic review. Hum Reprod Update 17: 620-627.
55. Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, et al. (2015) Somatic mutation in single human neurons tracks developmental and transcriptional history. Science 350: 94-98.
56. Shapiro E, Biezuner T, Linnarsson S (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 14: 618-630.
57. Amat F, Lemon W, Mossing DP, McDole K, Wan Y, et al. (2014) Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. Nat Meth advance online publication.
58. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, et al. (2011) Tumour evolution inferred by single-cell sequencing. Nature 472: 90-94.
59. Wang Y, Waters J, Leung ML, Unruh A, Roh W, et al. (2014) Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature 512: 155-160.
60. Bolognia JL, Orlow SJ, Glick SA (1994) Lines of Blaschko. Journal of the American Academy of Dermatology 31: 157-190.

61. Shirley MD, Tang HT, Gallione BA, Baugher JD, Frelin LP, et al. (2013) Sturge-Weber Syndrome and Port-Wine Stains Caused by Somatic Mutation in *GNAQ*. N Engl J Med 368: 1971-1979.
62. Van Raamsdonk CD, Griewank KG, Crosby MB, Garrido MC, Vemula S, et al. (2010) Mutations in GNA11 in uveal melanoma. N Engl J Med 363: 2191-2199.
63. Collins MT, Singer FR, Eugster E (2012) McCune-Albright syndrome and the extraskeletal manifestations of fibrous dysplasia. Orphanet J Rare Dis 7 Suppl 1: S4.
64. Bastepe M, Juppner H (2005) GNAS locus and pseudohypoparathyroidism. Horm Res 63: 65-74.
65. Poduri A, Evrony GD, Cai X, Elhosary PC, Beroukhim R, et al. (2012) Somatic activation of AKT3 causes hemispheric developmental brain malformations. Neuron 74: 41-48.
66. Lindhurst MJ, Sapp JC, Teer JK, Johnston JJ, Finn EM, et al. (2011) A mosaic activating mutation in AKT1 associated with the Proteus syndrome. N Engl J Med 365: 611-619.
67. Hussain K, Challis B, Rocha N, Payne F, Minic M, et al. (2011) An activating mutation of AKT2 and human hypoglycemia. Science 334: 474.
68. Jacobs KB, Yeager M, Zhou W, Wacholder S, Wang Z, et al. (2012) Detectable clonal mosaicism and its relationship to aging and cancer. Nat Genet 44: 651-658.
69. Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, et al. (2012) Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat Genet 44: 642-650.
70. Aghili L, Foo J, DeGregori J, De S (2014) Patterns of somatically acquired amplifications and deletions in apparently normal tissues of ovarian cancer patients. Cell Rep 7: 1310-1319.
71. Costa T, Lambert M, Teshima I, Ray PN, Richer CL, et al. (1998) Monozygotic twins with 45,X/46,XY mosaicism discordant for phenotypic sex. Am J Med Genet 75: 40-44.
72. Fujimoto A, Boelter WD, Sparkes RS, Lin MS, Battersby K (1991) Monozygotic twins of discordant sex both with 45, X/46, X, idic (Y) mosaicism. American journal of medical genetics 41: 239-245.
73. Zeng S, Patil SR, Yankowitz J (2003) Prenatal detection of mosaic trisomy 1q due to an unbalanced translocation in one fetus of a twin pregnancy following in vitro fertilization: A postzygotic error. American Journal of Medical Genetics Part A 120A: 464-469.
74. Kaplan L, Foster R, Shen Y, Parry DM, McMaster ML, et al. (2010) Monozygotic twins discordant for neurofibromatosis 1. American journal of medical genetics Part A 152: 601-606.
75. Helderman-van den Enden A, Maaswinkel-Mooij P, Hoogendoorn E, Willemsen R, Maat-Kievit J, et al. (1999) Monozygotic twin brothers with the fragile X syndrome: different CGG repeats and different mental capacities. Journal of medical genetics 36: 253-257.
76. Piotrowski A, Bruder CE, Andersson R, Diaz de Stahl T, Menzel U, et al. (2008) Somatic mosaicism for copy number variation in differentiated human tissues. Hum Mutat 29: 1118-1124.

77. O'Huallachain M, Karczewski KJ, Weissman SM, Urban AE, Snyder MP (2012) Extensive genetic variation in somatic human tissues. Proc Natl Acad Sci U S A 109: 18018-18023.
78. Abyzov A, Mariani J, Palejev D, Zhang Y, Haney MS, et al. (2012) Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature 492: 438-442.
79. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, et al. (2011) Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479: 534-537.
80. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, et al. (2009) L1 retrotransposition in human neural progenitor cells. Nature 460: 1127-1131.
81. Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, et al. (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435: 903-910.
82. Rehen SK, McConnell MJ, Kaushal D, Kingsbury MA, Yang AH, et al. (2001) Chromosomal variation in neurons of the developing and adult mammalian nervous system. Proc Natl Acad Sci U S A 98: 13361-13366.
83. Cai X, Evrony GD, Lehmann HS, Elhosary PC, Mehta BK, et al. (2014) Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. Cell Rep 8: 1280-1289.
84. Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, et al. (2012) Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell 151: 483-496.
85. McConnell MJ, Lindberg MR, Brennand KJ, Piper JC, Voet T, et al. (2013) Mosaic copy number variation in human neurons. Science 342: 632-637.
86. Kalousek DK, Dill FJ (1983) Chromosomal mosaicism confined to the placenta in human conceptions. Science 221: 665-667.
87. Taylor TH, Gitlin SA, Patrick JL, Crain JL, Wilson JM, et al. (2014) The origin, mechanisms, incidence and clinical consequences of chromosomal mosaicism in humans. Hum Reprod Update 20: 571-581.
88. Kalousek DK, Vekemans M (1996) Confined placental mosaicism. J Med Genet 33: 529-533.
89. Vanneste E, Voet T, Le Caignec C, Ampe M, Konings P, et al. (2009) Chromosome instability is common in human cleavage-stage embryos. Nat Med 15: 577-583.
90. Chen EZ, Chiu RW, Sun H, Akolekar R, Chan KC, et al. (2011) Noninvasive prenatal diagnosis of fetal trisomy 18 and trisomy 13 by maternal plasma DNA sequencing. PLoS One 6: e21791.
91. Ito Y, Tanaka F, Yamamoto M, Doyu M, Nagamatsu M, et al. (1998) Somatic Mosaicism of the Expanded CAG Trinucleotide Repeat in mRNAs for the Responsible Gene of Machado-Joseph Disease (MJD), Dentatorubral-Pallidoluysian Atrophy (DRPLA), and Spinal and Bulbar Muscular Atrophy (SBMA). Neurochemical Research 23: 25-32.
92. James CD, Carlbom E, Nordenskjold M, Collins VP, Cavenee WK (1989) Mitotic recombination of chromosome 17 in astrocytomas. Proceedings of the National Academy of Sciences 86: 2858-2862.

93. Kleczkowska A, Fryns JP, Van den Berghe H (1990) On the variable effect of mosaic normal/balanced chromosomal rearrangements in man. J Med Genet 27: 505-507.
94. Kotzot D, Schmitt S, Bernasconi F, Robinson WP, Lurie IW, et al. (1995) Uniparental disomy 7 in Silver—Russell syndrome and primordial growth retardation. Human Molecular Genetics 4: 583-587.
95. Rodríguez-Santiago B, Malats N, Rothman N, Armengol L, Garcia-Closas M, et al. (2010) Mosaic Uniparental Disomies and Aneuploidies as Large Structural Variants of the Human Genome. The American Journal of Human Genetics 87: 129-138.
96. Slatter RE, Elliott M, Welham K, Carrera M, Schofield PN, et al. (1994) Mosaic uniparental disomy in Beckwith-Wiedemann syndrome. Journal of Medical Genetics 31: 749-753.
97. Watson IR, Takahashi K, Futreal PA, Chin L (2013) Emerging patterns of somatic mutations in cancer. Nature Rev Genet 14: 703-718.
98. Youssoufian H, Pyeritz RE (2002) Mechanisms and consequences of somatic mosaicism in humans. Nat Rev Genet 3: 748-758.
99. Zori RT, Gray BA, Bent-Williams A, Driscoll DJ, Williams CA, et al. (1993) Preaxial acrofacial dysostosis (Nager syndrome) associated with an inherited and apparently balanced X;9 translocation: Prenatal and postnatal late replication studies. American Journal of Medical Genetics 46: 379-383.
100. Walsh C, Cepko CL (1992) Widespread dispersion of neuronal clones across functional regions of the cerebral cortex. Science 255: 434-440.
101. Pleasure SJ, Anderson S, Hevner R, Bagri A, Marin O, et al. (2000) Cell migration from the ganglionic eminences is required for the development of hippocampal GABAergic interneurons. Neuron 28: 727-740.
102. Hohn A, Leibrock J, Bailey K, Barde Y-A (1990) Identification and characterization of a novel member of the nerve growth factor/brain-derived neurotrophic factor family. Nature 344: 339-341.
103. Leibrock J, Lottspeich F, Hohn A, Hofer M, Hengerer B, et al. (1989) Molecular cloning and expression of brain-derived neurotrophic factor. Nature 341: 149-152.
104. Levi-Montalcini R (1964) GROWTH CONTROL OF NERVE CELLS BY A PROTEIN FACTOR AND ITS ANTISERUM: DISCOVERY OF THIS FACTOR MAY PROVIDE NEW LEADS TO UNDERSTANDING OF SOME NEUROGENETIC PROCESSES. Science 143: 105-110.
105. Kurek KC, Luks VL, Ayturk UM, Alomari AI, Fishman SJ, et al. (2012) Somatic mosaic activating mutations in PIK3CA cause CLOVES syndrome. Am J Hum Genet 90: 1108-1115.
106. Li H (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30: 2843-2851.
107. Gerstung M, Papaemmanuil E, Campbell PJ (2014) Subclonal variant calling with multiple samples and prior knowledge. Bioinformatics.
108. Crotwell PL, Hoyme HE (2012) Advances in whole-genome genetic testing: from chromosomes to microarrays. Curr Probl Pediatr Adolesc Health Care 42: 47-73.
109. Bushman DM, Chun J (2013) The genomically mosaic brain: aneuploidy and more in neural diversity and disease. Semin Cell Dev Biol 24: 357-369.
110. Notini AJ, Craig JM, White SJ (2008) Copy number variation and mosaicism. Cytogenet Genome Res 123: 270-277.

111. Vorsanova SG, Yurov YB, Iourov IY (2010) Human interphase chromosomes: a review of available molecular cytogenetic technologies. Mol Cytogenet 3: 1.
112. Imataka G, Arisaka O (2012) Chromosome analysis using spectral karyotyping (SKY). Cell Biochem Biophys 62: 13-17.
113. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, et al. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258: 818-821.
114. Pinkel D, Albertson DG (2005) Array comparative genomic hybridization and its applications in cancer. Nat Genet 37 Suppl: S11-17.
115. Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12: 363-376.
116. Bignell GR, Huang J, Greshock J, Watt S, Butler A, et al. (2004) High-Resolution Analysis of DNA Copy Number Using Oligonucleotide Microarrays. Genome Research 14: 287-295.
117. Mohr S, Leikauf GD, Keith G, Rihn BH (2002) Microarrays as Cancer Keys: An Array of Possibilities. Journal of Clinical Oncology 20: 3165-3175.
118. Baugher JD, Baugher BD, Shirley MD, Pevsner J (2013) Sensitive and specific detection of mosaic chromosomal abnormalities using the Parent-of-Origin-based Detection (POD) method. BMC Genomics 14: 367.
119. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11: 733-739.
120. Chen GK, Chang X, Curtis C, Wang K (2013) Precise inference of copy number alterations in tumor samples from SNP arrays. Bioinformatics 29: 2964-2970.
121. Li A, Liu Z, Lezon-Geyda K, Sarkar S, Lannin D, et al. (2011) GPHMM: an integrated hidden Markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays. Nucleic Acids Research 39: 4928-4941.
122. Liu Z, Li A, Schulz V, Chen M, Tuck D (2010) MixHMM: inferring copy number variation and allelic imbalance using SNP arrays and tumor samples mixed with stromal cells. PLoS One 5: e10909.
123. Rancoita PM, Hutter M, Bertoni F, Kwee I (2010) An integrated Bayesian analysis of LOH and copy number data. BMC bioinformatics 11: 1.
124. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38: 1767-1771.
125. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19: 1586-1592.
126. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, et al. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22: 568-576.
127. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272-276.

128. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, et al. (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proceedings of the National Academy of Sciences 106: 19096-19101.
129. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, et al. (2007) Genome-wide in situ exon capture for selective resequencing. Nat Genet 39: 1522-1527.
130. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, et al. (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28: 311-317.
131. Roth A, Ding J, Morin R, Crisan A, Ha G, et al. (2012) JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics 28: 907-913.
132. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, et al. (2012) Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28: 1811-1817.
133. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, et al. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31: 213-219.
134. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
135. Amarasinghe KC, Li J, Halgamuge SK (2013) CoNVEX: copy number variation estimation in exome sequencing data using HMM. BMC Bioinformatics 14 Suppl 2: S2.
136. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, et al. (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 28: 423-425.
137. Layer RM, Chiang C, Quinlan AR, Hall IM (2014) LUMPY: A probabilistic framework for structural variant discovery. Genome Biol 15: R84.
138. Chen M, Gunel M, Zhao H (2013) SomatiCA: identifying, characterizing and quantifying somatic copy number aberrations from cancer genome sequencing data. PLoS One 8: e78143.
139. Yadav VK, De S (2014) An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. Briefings in bioinformatics: bbu002.
140. Ding L, Wendl MC, McMichael JF, Raphael BJ (2014) Expanding the computational toolbox for mining cancer genomes. Nature Reviews Genetics 15: 556-570.
141. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, et al. (2002) Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A 99: 5261-5266.
142. Hosono S, Faruqi AF, Dean FB, Du Y, Sun Z, et al. (2003) Unbiased Whole-Genome Amplification Directly From Clinical Samples. Genome Research 13: 954-964.
143. Pugh TJ, Delaney AD, Farnoud N, Flibotte S, Griffith M, et al. (2008) Impact of whole genome amplification on analysis of copy number variants. Nucleic Acids Res 36: e80.
144. Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, et al. (2012) Genome-wide copy number analysis of single cells. Nat Protoc 7: 1024-1041.

145. Gundry M, Li W, Maqbool SB, Vijg J (2012) Direct, genome-wide assessment of DNA mutations in single cells. Nucleic Acids Res 40: 2032-2040.
146. Hou Y, Song L, Zhu P, Zhang B, Tao Y, et al. (2012) Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell 148: 873-885.
147. Xu X, Hou Y, Yin X, Bao L, Tang A, et al. (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell 148: 886-895.
148. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, et al. (2006) The consensus coding sequences of human breast and colorectal cancers. Science 314: 268-274.
149. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, et al. (2012) The landscape of cancer genes and mutational processes in breast cancer. Nature 486: 400-404.
150. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458: 719-724.
151. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144: 646-674.
152. McCubrey JA, Steelman LS, Chappell WH, Abrams SL, Montalto G, et al. (2012) Mutations and deregulation of Ras/Raf/MEK/ERK and PI3K/PTEN/Akt/mTOR cascades which alter therapy response. Oncotarget 3: 954-987.
153. Downward J (2003) Targeting RAS signalling pathways in cancer therapy. Nat Rev Cancer 3: 11-22.
154. Liaw D, Marsh DJ, Li J, Dahia PLM, Wang SI, et al. (1997) Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. Nat Genet 16: 64-67.
155. Malkin D, Li F, Strong L, Fraumeni J, Nelson C, et al. (1990) Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science 250: 1233-1238.
156. Morin PJ, Sparks AB, Korinek V, Barker N, Clevers H, et al. (1997) Activation of β-Catenin-Tcf Signaling in Colon Cancer by Mutations in β-Catenin or APC. Science 275: 1787-1790.
157. Cleaver JE (1968) Defective Repair Replication of DNA in Xeroderma Pigmentosum. Nature 218: 652-656.
158. Miyaki M, Konishi M, Tanaka K, Kikuchi-Yanoshita R, Muraoka M, et al. (1997) Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. Nat Genet 17: 271-272.
159. Moynahan ME, Chiu JW, Koller BH, Jasin M (1999) Brca1 Controls Homology-Directed DNA Repair. Molecular Cell 4: 511-518.
160. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, et al. (1995) Identification of the breast cancer susceptibility gene BRCA2. Nature 378: 789-792.
161. Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, et al. (2012) Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell 150: 1121-1134.
162. Collins FS, Barker AD (2007) Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. Sci Am 296: 50-57.

163. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, et al. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 39: D945-950.
164. Debeljak M, Freed DN, Welch JA, Haley L, Beierl K, et al. (2014) Haplotype Counting by Next-Generation Sequencing for Ultrasensitive Human DNA Detection. The Journal of Molecular Diagnostics 16: 495-503.
165. Armitage P, Doll R (1954) The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis. Br J Cancer 8: 1-12.
166. Curtis HJ (1963) Biological mechanisms underlying the aging process. Science 141: 686-694.
167. Szilard L (1959) ON THE NATURE OF THE AGING PROCESS. Proc Natl Acad Sci U S A 45: 30-45.
168. Albertson TM, Ogawa M, Bugni JM, Hays LE, Chen Y, et al. (2009) DNA polymerase ε and δ proofreading suppress discrete mutator and cancer phenotypes in mice. Proceedings of the National Academy of Sciences 106: 17101-17104.
169. Goldsby RE, Hays LE, Chen X, Olmsted EA, Slayton WB, et al. (2002) High incidence of epithelial cancers in mice deficient for DNA polymerase δ proofreading. Proceedings of the National Academy of Sciences 99: 15560-15565.
170. Goldsby RE, Lawrence NA, Hays LE, Olmsted EA, Chen X, et al. (2001) Defective DNA polymerase-[delta] proofreading causes cancer susceptibility in mice. Nat Med 7: 638-639.
171. Trifunovic A, Wredenberg A, Falkenberg M, Spelbrink JN, Rovio AT, et al. (2004) Premature ageing in mice expressing defective mitochondrial DNA polymerase. Nature 429: 417-423.
172. Vermulst M, Bielas JH, Kujoth GC, Ladiges WC, Rabinovitch PS, et al. (2007) Mitochondrial point mutations do not limit the natural lifespan of mice. Nat Genet 39: 540-543.
173. Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JHJ (2014) Understanding nucleotide excision repair and its roles in cancer and ageing. Nat Rev Mol Cell Biol 15: 465-481.
174. Mohaghegh P, Hickson ID (2001) DNA helicase deficiencies associated with cancer predisposition and premature ageing disorders. Human Molecular Genetics 10: 741-746.
175. Hanks S, Coleman K, Reid S, Plaja A, Firth H, et al. (2004) Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in BUB1B. Nat Genet 36: 1159-1161.
176. Date H, Onodera O, Tanaka H, Iwabuchi K, Uekawa K, et al. (2001) Early-onset ataxia with ocular motor apraxia and hypoalbuminemia is caused by mutations in a new HIT superfamily gene. Nat Genet 29: 184-188.
177. Monnat Jr RJ (2010) Human RECQ helicases: Roles in DNA metabolism, mutagenesis and cancer biology. Seminars in Cancer Biology 20: 329-339.
178. Moreira M-C, Barbot C, Tachi N, Kozuka N, Uchida E, et al. (2001) The gene mutated in ataxia-ocular apraxia 1 encodes the new HIT/Zn-finger protein aprataxin. Nat Genet 29: 189-193.
179. Niedernhofer LJ (2008) Tissue-specific accelerated aging in nucleotide excision repair deficiency. Mechanisms of Ageing and Development 129: 408-415.

180. Burdick D, Soreghan B, Kwon M, Kosmoski J, Knauer M, et al. (1992) Assembly and aggregation properties of synthetic Alzheimer's A4/beta amyloid peptide analogs. Journal of Biological Chemistry 267: 546-554.
181. Goldfarb LG, Brown P, McCombie WR, Goldgaber D, Swergold GD, et al. (1991) Transmissible familial Creutzfeldt-Jakob disease associated with five, seven, and eight extra octapeptide coding repeats in the PRNP gene. Proceedings of the National Academy of Sciences 88: 10926-10930.
182. Alzualde A, Moreno F, Martínez-Lage P, Ferrer I, Gorostidi A, et al. (2010) Somatic mosaicism in a case of apparently sporadic Creutzfeldt-Jakob disease carrying a de novo D178N mutation in the PRNP gene. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 153B: 1283-1291.
183. Beck JA, Poulter M, Campbell TA, Uphill JB, Adamson G, et al. (2004) Somatic and germline mosaicism in sporadic early-onset Alzheimer's disease. Human Molecular Genetics 13: 1219-1224.
184. Eikelenboom P, Bate C, Van Gool WA, Hoozemans JJ, Rozemuller JM, et al. (2002) Neuroinflammation in Alzheimer's disease and prion disease. Glia 40: 232-239.
185. Kane MD, Lipinski WJ, Callahan MJ, Bian F, Durham RA, et al. (2000) Evidence for seeding of beta -amyloid by intracerebral infusion of Alzheimer brain extracts in beta -amyloid precursor protein-transgenic mice. J Neurosci 20: 3606-3611.
186. Meyer-Luehmann M, Coomaraswamy J, Bolmont T, Kaeser S, Schaefer C, et al. (2006) Exogenous induction of cerebral beta-amyloidogenesis is governed by agent and host. Science 313: 1781-1784.
187. Lu J-X, Qiang W, Yau W-M, Schwieters Charles D, Meredith Stephen C, et al. (2013) Molecular Structure of β-Amyloid Fibrils in Alzheimer's Disease Brain Tissue. Cell 154: 1257-1268.
188. Stöhr J, Condello C, Watts JC, Bloch L, Oehler A, et al. (2014) Distinct synthetic Aβ prion strains producing different amyloid deposits in bigenic mice. Proceedings of the National Academy of Sciences 111: 10329-10334.
189. Watts JC, Condello C, Stöhr J, Oehler A, Lee J, et al. (2014) Serial propagation of distinct strains of Aβ prions from Alzheimer's disease patients. Proceedings of the National Academy of Sciences 111: 10323-10328.
190. Ott A, Breteler MM, van Harskamp F, Claus JJ, van der Cammen TJ, et al. (1995) Prevalence of Alzheimer's disease and vascular dementia: association with education. The Rotterdam study. BMJ 310: 970-973.
191. Chapman J, Ben-Israel J, Goldhammer Y, Korczyn AD (1994) The risk of developing Creutzfeldt-Jakob disease in subjects with the PRNP gene codon 200 point mutation. Neurology 44: 1683-1686.
192. Mirzaa G, Conaway R, Graham JM, Jr., Dobyns WB (2013) PIK3CA-Related Segmental Overgrowth. In: Pagon RA, Adam MP, Ardinger HH, al. E, editors. GeneReviews. Seattle (WA): University of Washington, Seattle.
193. Samuels Y, Wang Z, Bardelli A, Silliman N, Ptak J, et al. (2004) High frequency of mutations of the PIK3CA gene in human cancers. Science 304: 554.
194. Carpten JD, Faber AL, Horn C, Donoho GP, Briggs SL, et al. (2007) A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. Nature 448: 439-444.

195. Riviere JB, Mirzaa GM, O'Roak BJ, Beddaoui M, Alcantara D, et al. (2012) De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. Nat Genet 44: 934-940.
196. Lee JH, Huynh M, Silhavy JL, Kim S, Dixon-Salazar T, et al. (2012) De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nat Genet 44: 941-945.
197. Zhang Y, Gao X, Saucedo LJ, Ru B, Edgar BA, et al. (2003) Rheb is a direct target of the tuberous sclerosis tumour suppressor proteins. Nat Cell Biol 5: 578-581.
198. Curatolo P, Bombardieri R, Jozwiak S Tuberous sclerosis. The Lancet 372: 657-668.
199. Henske EP, Wessner LL, Golden J, Scheithauer BW, Vortmeyer AO, et al. (1997) Loss of tuberin in both subependymal giant cell astrocytomas and angiomyolipomas supports a two-hit model for the pathogenesis of tuberous sclerosis tumors. Am J Pathol 151: 1639-1647.
200. Tsang E, Birch P, Friedman JM (2012) Valuing gene testing in children with possible neurofibromatosis 1. Clin Genet 82: 591-593.
201. Pansuriya TC, van Eijk R, d'Adamo P, van Ruler MA, Kuijjer ML, et al. (2011) Somatic mosaic IDH1 and IDH2 mutations are associated with enchondroma and spindle cell hemangioma in Ollier disease and Maffucci syndrome. Nat Genet 43: 1256-1261.
202. Jamuar SS, Lam A-TN, Kircher M, D'Gama AM, Wang J, et al. (2014) Somatic Mutations in Cerebral Cortical Malformations. New England Journal of Medicine 371: 733-743.
203. Choate KA, Lu Y, Zhou J, Choi M, Elias PM, et al. (2010) Mitotic recombination in patients with ichthyosis causes reversion of dominant mutations in KRT10. Science 330: 94-97.
204. Pasmooij AM, Jonkman MF, Uitto J (2012) Revertant mosaicism in heritable skin diseases: mechanisms of natural gene therapy. Discov Med 14: 167-179.
205. Pasmooij AM, Pas HH, Bolling MC, Jonkman MF (2007) Revertant mosaicism in junctional epidermolysis bullosa due to multiple correcting second-site mutations in LAMB3. J Clin Invest 117: 1240-1248.
206. Hirschhorn R, Yang DR, Puck JM, Huie ML, Jiang CK, et al. (1996) Spontaneous in vivo reversion to normal of an inherited mutation in a patient with adenosine deaminase deficiency. Nat Genet 13: 290-295.
207. Soulier J, Leblanc T, Larghero J, Dastot H, Shimamura A, et al. (2005) Detection of somatic mosaicism and classification of Fanconi anemia patients by analysis of the FA/BRCA pathway. Blood 105: 1329-1336.
208. De S (2011) Somatic mosaicism in healthy human tissues. Trends Genet 27: 217-223.
209. Insel TR (2014) Brain somatic mutations: the dark matter of psychiatric genetics? Mol Psychiatry 19: 156-158.
210. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, et al. (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515: 209-215.
211. Gaugler T, Klei L, Sanders SJ, Bodea CA, Goldberg AP, et al. (2014) Most genetic risk for autism resides with common variation. Nat Genet 46: 881-885.
212. Iossifov I, O/'Roak BJ, Sanders SJ, Ronemus M, Krumm N, et al. (2014) The contribution of de novo coding mutations to autism spectrum disorder. Nature 515: 216-221.

213. Krumm N, O'Roak BJ, Shendure J, Eichler EE (2014) A de novo convergence of autism genetics and molecular neuroscience. Trends Neurosci 37: 95-105.
214. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.
215. Stoner R, Chow ML, Boyle MP, Sunkin SM, Mouton PR, et al. (2014) Patches of disorganization in the neocortex of children with autism. N Engl J Med 370: 1209-1219.
216. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al. (2007) Strong association of de novo copy number mutations with autism. Science 316: 445-449.
217. Zhao X, Leotta A, Kustanovich V, Lajonchere C, Geschwind DH, et al. (2007) A unified genetic theory for sporadic and inherited autism. Proceedings of the National Academy of Sciences 104: 12831-12836.
218. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, et al. (2008) Structural variation of chromosomes in autism spectrum disorder. American Journal of Human Genetics 82: 477-488.
219. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, et al. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466: 368-372.
220. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, et al. (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet 43: 585-589.
221. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, et al. (2011) Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 70: 863-885.
222. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, et al. (2012) De novo gene disruptions in children on the autistic spectrum. Neuron 74: 285-299.
223. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485: 242-245.
224. O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, et al. (2012) Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science 338: 1619-1622.
225. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, et al. (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485: 246-250.
226. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, et al. (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485: 237-241.
227. Ronemus M, Iossifov I, Levy D, Wigler M (2014) The role of de novo mutations in the genetics of autism spectrum disorders. Nat Rev Genet 15: 133-141.
228. Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, et al. (2015) Excess of rare, inherited truncating mutations in autism. Nat Genet 47: 582-588.
229. Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, et al. (2014) Genome sequencing identifies major causes of severe intellectual disability. Nature 511: 344-347.

230. O'Roak BJ, Stessman HA, Boyle EA, Witherspoon KT, Martin B, et al. (2014) Recurrent de novo mutations implicate novel genes underlying simplex autism risk. Nat Commun 5.
231. Dal GM, Erguner B, Sagiroglu MS, Yuksel B, Onat OE, et al. (2014) Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. J Med Genet 51: 455-459.
232. Li R, Montpetit A, Rousseau M, Wu SY, Greenwood CM, et al. (2014) Somatic point mutations occurring early in development: a monozygotic twin study. J Med Genet 51: 28-34.
233. Acuna-Hidalgo R, Bo T, Kwint Michael P, van de Vorst M, Pinelli M, et al. (2015) Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. The American Journal of Human Genetics 97: 67-74.
234. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
235. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.
236. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.
237. Koster J, Rahmann S (2012) Snakemake--a scalable bioinformatics workflow engine. Bioinformatics 28: 2520-2522.
238. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6: 80-92.
239. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.
240. Siepel A, Pollard KS, Haussler D (2006) New Methods for Detecting Lineage-Specific Selection. In: Apostolico A, Guerra C, Istrail S, Pevzner PA, Waterman M, editors. Research in Computational Molecular Biology: 10th Annual International Conference, RECOMB 2006, Venice, Italy, April 2-5, 2006 Proceedings. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 190-205.
241. Coordinators NR, Acland A, Agarwala R, Barrett T, Beck J, et al. (2014) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 42: D7-D17.
242. Lek M, Karczewski K, Minikel E, Samocha K, Banks E, et al. (2015) Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv.
243. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422-1423.
244. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, et al. (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 13: 134.
245. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The Human Genome Browser at UCSC. Genome Research 12: 996-1006.

246. Holstege H, Pfeiffer W, Sie D, Hulsman M, Nicholas TJ, et al. (2014) Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. Genome Research 24: 733-742.
247. Genovese G, Kahler AK, Handsaker RE, Lindberg J, Rose SA, et al. (2014) Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N Engl J Med 371: 2477-2487.
248. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, et al. (2014) Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotech 32: 246-251.
249. Fischbach GD, Lord C The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. Neuron 68: 192-195.
250. Yadav VK, DeGregori J, De S (2016) The landscape of somatic mutations in protein coding genes in apparently benign human tissues carries signatures of relaxed purifying selection. Nucleic Acids Res 44: 2075-2084.

CURRICULUM VITAE

Donald Freed 12-14-2016

Educational History:

Ph.D. expected	2016	Program in Neuroscience Mentor: Jonathan Pevsner	Johns Hopkins School of Medicine
B.S.		Biochemistry	Hillsdale College

Other Professional Experience:

Research rotation	2013-2013	Lab of Dr. Akhilesh Pandey, Johns Hopkins SOM
Research rotation	2013-2013	Lab of Dr. Mario Amzel, Johns Hopkins SOM
Research rotation	2012-2012	Lab of Dr. Jonathan Pevsner, Johns Hopkins SOM
Research rotation	2012-2012	Lab of Dr. Heng Zhu, Johns Hopkins SOM
Undergraduate Researcher	2011-2012	Lab of Dr. Christopher Hamilton, Hillsdale College

Honors:

May 2012	Departmental Honors	Hillsdale College
May 2012	Magna cum Laude	Hillsdale College
April 2012	Summer Research Award	Hillsdale College

Publications:

Freed D, *et al.* (in preparation) An assessment of genetic variation in 22 cases of self-injurious behavior from whole-genome sequence data.

McConnell M, ***et al.*** (submitted) Intersection of Diverse Neuronal Genomes and Neurological Disease: The Brain Somatic Mosaicism Network.

Freed D, Pevsner J. (2016) The Contribution of Mosaic Variants to Autism Spectrum Disorder. PLoS Genet, vol. 12; e1006245.

Freed D, Stevens EL, Pevsner J. (2014) Somatic mosaicism in the human genome. Genes, vol. 5; 1064-1094.

Debeljak M, **Freed DN**, Welch JA, *et al.* (2014) Haplotype counting by next-generation sequencing for ultrasensitive human DNA detection. J Mol Diagn, vol. 16; 495-503.

Kim MS, Pinto S, Getnet D, ***et al***. (2014) A draft map of the human proteome. Nature, vol. 509; 575–581

Presentations:

Freed D, Pevsner J (2016) Contribution of mosaic variation to autism spectrum disorders. American Society of Human Genetics Meeting, Vancouver, BC, October 2016

Freed D, Pevsner J (2016) Mosaic variation in autism and bipolar disorder: recent findings and novel methods. Brain Somatic Mosaicism Network Workshop, Baltimore, MD, September 2016

Freed D, Pevsner J (2015) Deleterious mosaic variants among *de novo* mutations in autism and bipolar disorder, Brain Somatic Mosaicism Network Workshop, Rockville, MD, November 2015

Posters:

Freed D, Pevsner J (2015) Somatic Mosaicism in Autism Spectrum Disorder. Symposium and Poster Session on Genomics and Bioinformatics, Baltimore, MD, October 2015

Freed D, Pevsner J (2015) Somatic Mosaicism in Autism Spectrum Disorder. BCMB Annual Retreat, Harbourtowne, MD, October 2015